Methodological Approach to Conformational Search. A Study Case: Cyclodextrins

Kepa Koldo Burusco Goñi

PhD Thesis

Doctoral Program in Chemistry

Supervised by Professor Dr. Carlos Jaime Cardiel

Department of Chemistry

Faculty of Sciences

2009

Departamento de Química

Facultad de Ciencias

Memoria presentada para aspirar al Grado de Doctor por Kepa Koldo Burusco Goñi.

Leído y conforme:

Dr. Carlos Jaime Cardiel **Kepa Koldo Burusco Goñi**

Bellaterra, a ____ de _____ de 2009.

Agradecimientos (Acknowledgments)

La investigación llevada a cabo en esta Tesis Doctoral y recogida en la presente memoria ha sido posible gracias a la financiación de las siguientes entidades públicas y privadas:

- Proyecto: "Avances en el reconocimiento quiral. Nuevos disolventes de solvatación quiral y asociación supramolecular". PPQ2000-0369.
- Proyecto: "Compuestos difuncionales y macrocíclicos con propiedades enantiodiferenciadoras. Preparación y estudio estructural, teórico y experimental." BQU2003-01231.
- Proyecto: "Compuestos y materiales polifuncionales y/o macrocíclicos con cavidades enantiodiferenciadoras. Preparación y estudio estructural, teórico y experimental." Ministerio de Educación y Ciencia (MEC)
- Proyecto: "Estudio computacional de propiedades macroscópicas de macromoléculas de interés industrial." En colaboración con Industrial Química Lasem dentro del proyecto PIIBE (Proyecto de Investigación para el Impulso del Biodiésel en España), enmarcado dentro de los proyectos CENIT (Consorcio Estratégico Nacional de Investigación Técnica).
- Proyecto: "Dinámica Molecular de los complejos de inclusión: Ciclodextrinas y Rotaxanos." CESCA - Generalitat de Cataluña – CICyT.

Asimismo, quiero mostrar mi agradecimiento personal por las ayudas económicas recibidas de las siguientes instituciones:

- Beca de Formación del Profesorado Universitario (FPU) del Ministerio de Educación y Ciencia.
- Beca de colaboración de Industrial Química Lasem.
- Contratos de Profesor Asociado de la Universidad Autónoma de Barcelona.

Agradecimientos Personales

Estos años han marcado mi vida. No todo han sido alegrías –mentiría si dijera lo contrario- pero también ha habido buenos momentos que no olvidaré. El doctorado, además de un reto académico, ha sido también un largo viaje donde uno se acaba descubriendo a sí mismo tal y como es; conociendo lo mejor y lo peor que lleva dentro. Hay ciertas personas que en este tiempo han aportado su ayuda para que yo pudiera completar este viaje y quisiera hacerlas partícipes de mi gratitud ahora en estas líneas. Sin su apoyo, estas páginas no habrían sido posibles.

Comenzaré dando las gracias a mis padres, ambos trabajadores incansables, por haberme apoyado en esta empresa incierta.

Al Dr. Carlos Jaime Cardiel, por haberme aceptado en su grupo y permitirme continuar en la UAB para acabar esta Tesis Doctoral a pesar de todo el tiempo que anduve perdido y sin encontrar el camino.

Quiero agradecer especialmente por el esfuerzo realizado al Dr. Francesc Sánchez Ferrando quien, sin ser mi director de tesis ni tampoco un especialista en la materia, ha tenido la amabilidad de leer el manuscrito íntegro y revisar la redacción en inglés, así como aportar valiosos comentarios acerca de la ciencia escrita. Cualquier error que todavía permanezca en el texto es, evidentemente, de mi exclusiva responsabilidad.

Hago extensible mi gratitud al Dr. Albert Virgili por su cercanía y por invitarme a sus celebraciones en torno a una paella; y al Dr. Pere de March –recientemente incorporado a nuestro grupo- por el grado de implicación, la experiencia y la profesionalidad que ha aportado, así como por su trato exquisito.

He visto pasar a mucha gente por el laboratorio de quienes guardo agradables recuerdos: Martina, Pau y Miriam allá por los lejanos comienzos; Raquel y Marta más recientemente; y naturalmente, Don Javier, quien ha sido un excelente compañero de viaje y un maestro en los oscuros caminos del manual de AMBER. No puedo olvidarme de Verónica, David "el Sevillano" y Jacek; siendo un placer disfrutar de su compañía aquí durante sus estancias doctorales o postdoctorales. Mención especial merecen Sergio, Julen, Eva y Lisi; ya que gracias a su apoyo e insistencia comencé a escribir esta Tesis aun sin estar del todo convencido de tener material suficiente. Ahora, volviendo la vista atrás, sólo puedo decir que tenían razón. Muchas gracias a todos. Para terminar, a "los nuevos", Marta y Josep, sólo desearles mucho ánimo y que no desesperen cuando vengan tiempos difíciles.

Aunque últimamente nos hayamos ido distanciando en mayor o menor grado por diferentes razones, quiero agradecer por los buenos ratos que hemos compartido a Xavi Bagán, Sergio Marín, Manu, Xavi Palmer, Francisco, Adu, Anabel & Arnau, Toni y Marta Ávila. A todos ellos les deseo lo mejor en el futuro.

He de hacer en este punto mención especial mis amigos de Pamplona. Por una parte a la gente de Maristas: Óscar, Miguel Ángel y especialmente a la cuadrilla: Sergio, Iñigo, Alberto, y Eduardo. Por haber logrado con éxito que no perdiéramos el contacto en todo este tiempo y hacerme sentir como en casa incluso en la distancia. Por otra parte a mis compañeros de la Universidad de Navarra: Josean, Ysa, Miguel "Chopi", Jesús & Sonia y "Xevi" & Asun. Por las excursiones que terminaban alrededor de una buena comida; por las partidas al mus junto a unas bravas; por los días de playa en San Sebastián; y por los viajes juntos. A todos ellos un fuerte abrazo desde aquí.

No quiero dejar pasar la oportunidad de saludar particularmente a tres personas: Igor, Pedro y Miguel; por todos los buenos ratos en Mendigorría, Pamplona o Torla; por las visitas a BCN; y por todas las discusiones filosóficas que, aunque no han arreglado el mundo, nos han servido de excusa para juntarnos en el Café Vienés o en el Australiano un buen número de veces. Espero que en el futuro podamos retomar la conversación allá donde la dejamos: "Como decíamos ayer..."

Quiero igualmente dar las gracias a mis compañeros de piso a lo largo de estos años por crear el ambiente adecuado para poder seguir adelante. De mis años en Cerdanyola, a Miguel (también amigo de Pamplona y compañero de laboratorio en las quijotescas cruzadas contra AMBER), Iván y Adrià; por las amenas discusiones sobre matemáticas, termodinámica, mecánica cuántica o informática; y, como contrapunto, a Simona; por su cordura, pragmatismo y sensatez. De mi época actual en Barcelona, a Pedro (igualmente amigo de Maristas en Pamplona) y Javier; por las sesiones maratonianas de ajedrez y la agradable convivencia.

Esta enumeración no estaría completa si no incluyera en ella a Raúl & Delia, con quienes he disfrutado en compañía de Sergio –"el Señor Marqués"- de exquisitas especialidades mexicanas y catalanas, así como de excursiones, vacaciones y agradables fines de semana. Desde aquí les envío un afectuoso abrazo.

Antes de acabar, quisiera expresar mi agradecimiento a aquellos con quienes di mis primeros pasos en el mundo de la Ciencia, allá en el Departamento de Física y Matemática Aplicada de la Universidad de Navarra: el Dr. Diego Maza Ozcoidi y el ya tristemente desaparecido Dr. Carlos Pérez García.

Por último y definitivamente: a Woody, por su agradable y felina compañía mientras escribía con no poco esfuerzo esta memoria en casa. Y por supuesto; a los Grandes Maestros –cuyas obras he podido escuchar en el Auditorio y el Gran Teatro del Liceo- por llenar de sentido tantos momentos vacíos.

A todos ellos, mi sincero reconocimiento.

I. Sobre el Liderazgo, la Estructura de Mando y la Disciplina.

El libro de Sun Tzu (500 a.C.), *El Arte de la Guerra*, consiguió una gran aceptación por parte del Rey de Wu, que dijo, "He leído detenidamente tus trece capítulos, ¿puedo someter tu teoría de dirigir a los soldados a un pequeño examen?"

Sun Tzu respondió: "Señor, puede hacerlo".

El Rey de Wu preguntó: "¿Es posible hacer este examen a las mujeres?".

Sun Tzu respondió que sí, así que se hizo lo necesario para reunir 180 mujeres del Palacio. Sun Tzu las dividió en dos compañías con una de las concubinas favoritas del Rey al frente de cada una de ellas. Entonces les hizo coger unas lanzas y les dijo: "Supongo que sabéis la diferencia entre delante y detrás, derecha e izquierda".

Las Mujeres respondieron: "Sí".

Sun Tzu continuó: "Cuando, al sonido de los tambores, ordene 'mirada al frente', mirad hacia adelante. Cuando mande 'giro a la izquierda', virad hacia vuestra izquierda. Cuando ordene 'giro a la derecha', hacedlo hacia vuestra derecha. Cuando ordene 'media vuelta', girad hacia atrás".

Una vez explicadas todas las órdenes de mando, las mujeres afirmaron que lo habían entendido y tomaron lanzas para poder empezar la instrucción. Con el sonido de los tambores, Sun Tzu ordenó: "Giro a la derecha". En respuesta, las mujeres se echaron a reír.

Con gran paciencia, Sun Tzu advirtió: "Si las instrucciones y palabras de mando no están claras, si las órdenes no se entienden bien, hay que echarle la culpa al general". Entonces repitió las explicaciones varias veces. Finalmente ordenó que los tambores señalaran "giro a la izquierda" y de nuevo las mujeres se echaron a reír.

Entonces Sun Tzu dijo: "Si las instrucciones y palabras de mando no están claras, si las órdenes no se entienden bien, hay que echarle la culpa al general. Pero si las órdenes están claras y los soldados desobedecen, entonces la falta es de los oficiales". Inmediatamente ordenó que las mujeres que estaban al frente de las dos compañías fueran decapitadas.

Evidentemente, el Rey lo presenció todo desde un podio elevado y, al ver que sus dos concubinas preferidas estaban a punto de ser ejecutadas, se sintió alarmado y rápidamente envió un mensaje: "Ahora estamos bastante satisfechos de la habilidad del general para dirigir a sus tropas. Sin estas concubinas, mi comida y mi bebida no me sabrían bien. Es deseo del Rey que no sean ejecutadas".

Sun Tzu replicó: "Habiendo recibido la comisión del soberano de hacerme cargo y dirigir estas tropas, hay ciertas órdenes que no puedo aceptar". Inmediatamente mandó que las dos concubinas fueran decapitadas, para dar ejemplo, y eligió a las dos siguientes de la formación como nuevas líderes.

Los tambores empezaron a sonar y comenzó la instrucción. Las mujeres ejecutaron las maniobras exactamente como les habían ordenado, girando a la derecha o a la izquierda, marchando hacia adelante, dando la vuelta, arrodillándose o poniéndose en pie. La instrucción fue perfecta en precisión y no pronunciaron ni un solo sonido.

Sun Tzu envió un mensajero al Rey de Wu diciendo: "Su Majestad, los soldados están correctamente entrenados y perfectamente disciplinados. Están listos para su inspección. Hágales hacer lo que usted quiera. Como soberano, puede ordenarles que se lancen contra el fuego o el agua y no le desobedecerán".

El Rey respondió: "Nuestro comandante debería cesar la instrucción y volver a su campamento. No queremos bajar e inspeccionar las tropas".

Con gran calma, Sun Tzu sentenció: "A este Rey sólo le gustan las palabras y no es capaz de ponerlas en acción".

Gerald Michaelson & Steven Michaelson.
Sun Tzu para el Éxito.
El Arte de la Guerra (500 a.C.)
La Lección de las Concubinas.
Ediciones Deusto (2004).
Planeta DeAgostini Profesional y Formación, S.L.
ISBN: 84-234-2141-4.

*****

II. Sobre la Sociedad "Actual".

La misión del llamado "intelectual" es, en cierto modo, opuesta a la del político. La obra intelectual aspira, con frecuencia en vano, a aclarar un poco las cosas, mientras que la del político suele, por el contrario, consistir en confundirlas más de lo que estaban. Ser de la izquierda es, como ser de la derecha, una de las infinitas maneras que el hombre puede elegir para ser un imbécil: ambas, en efecto, son formas de la hemiplejía moral. Además, la persistencia de estos calificativos contribuye no poco a falsificar más aún la "realidad" del presente, ya falsa de por sí, porque se ha rizado el rizo de las experiencias políticas a que responden, como lo demuestra el hecho de que hoy las derechas prometen revoluciones y las izquierdas proponen tiranías. [...]

El politicismo integral, la absorción de todas las cosas y de todo el hombre por la política, es una y misma cosa con el fenómeno de rebelión de las masas que aquí se describe. La masa en rebeldía ha perdido toda capacidad [de religión y] de conocimiento. No puede tener dentro más que política, una política exorbitada, frenética, fuera de sí, puesto que pretende suplantar al conocimiento, [a la religión,] a la *sagesse*[1] — en fin, a las únicas cosas que por su sustancia son aptas para ocupar el centro de la mente humana. La política vacía al hombre de soledad e intimidad, y por eso es la predicación del politicismo integral una de las técnicas que se usan para socializarlo.

Cuando alguien nos pregunta qué somos en política o, anticipándose con la insolencia que pertenece al estilo de nuestro tiempo, nos adscribe a una, en vez de responder, debemos preguntar al impertinente qué piensa él que es el hombre y la naturaleza y la historia, qué es la sociedad y el individuo, la colectividad, el Estado, el uso, el derecho. La política se apresura a apagar las luces para que todos estos gatos resulten pardos.

Es preciso que el pensamiento europeo proporcione sobre todos estos temas nueva claridad. Para eso está ahí, no para hacer la rueda del pavo real en las reuniones académicas. Y es preciso que lo haga pronto, o, como Dante decía, que encuentre la salida:

... *studiate il passo,*
mentre che l'Occidente non s'annera[2].
(Purg., XXVII, 62 — 63.)

José Ortega y Gasset.
La Rebelión de las Masas (1930)
Pág. 32-33.
XX Edición (2006).
Revista de Occidente, S.A.
Alianza Editorial.
ISBN: 84-206-4101-4.

*****

III. Sobre uno mismo.

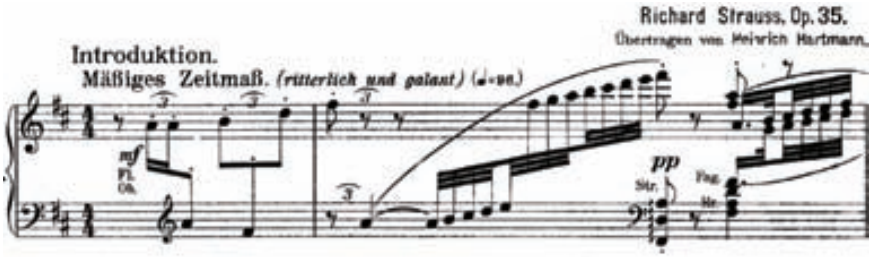
[Pero] el peor enemigo con que puedes encontrarte serás siempre tú mismo; a ti mismo te acechas tú en las cavernas y en los bosques.

Friedrich Nietzsche.
Así habló Zarathustra (1883-1885).
1ª Parte: Los Discursos de Zarathustra.
"Del camino del Creador".
Pág. 107.
I Edición Revisada (1997)
X Reimpresión (2007)
Alianza Editorial.
ISBN: 978-84-206-3319-0.

*****

[1] Sabiduría, prudencia, cordura.

[2] ...acelerad el paso,
mientras que el occidente no se oscurezca.



Richard Strauss
Don Quixote op. 35
(Fantastic Variations on a Theme of Knightly Character)
Tone Poem for Cello & Large Orchestra
(1898)
Introduction (Don Quixote's Leitmotiv)

"Science is a wonderful thing
if one does not have to earn one's living at it."

Albert Einstein

1 OBJECTIVES 3

2 INTRODUCTION: CONFORMATIONAL IMPORTANCE IN CHEMISTRY 9

2.1 CHEMISTRY BEYOND THE REACTIONS 11
2.1.1 "In the beginning God created Chirality…" 11
2.1.2 Brief History of Chirality: "The Birth of Stereochemistry" 12
2.1.3 Stereochemistry today 15
2.1.4 Supramolecular Chemistry: "Architecture in the Microscopic World" 16
2.1.4.1 Molecular Self-Assembly 16
2.1.4.2 Folding: Foldamers and Proteins 17
2.1.4.2.1 Foldamers 17
2.1.4.2.2 Proteins 18
2.1.4.2.3 Folding 19
2.1.4.3 Molecular Recognition 21
2.1.4.4 Host-Guest Chemistry 22
2.1.4.5 Mechanically-Interlocked Molecular Architectures 23
2.2 STEREOCHEMISTRY: IMPORTANCE OF THE MOLECULAR SPATIAL ARRANGEMENT 25
2.2.1 Structure-Activity Relationships I: Chirality 25
2.2.1.1 Specific biological activity for a single enantiomer 26
2.2.1.2 Identical qualitative and quantitative activity for any enantiomer 26
2.2.1.3 Identical qualitative - different quantitative activity for each enantiomer 27
2.2.1.4 Different qualitative activities for each enantiomer 28
2.2.1.5 The Thalidomide Disaster 29
2.2.1.6 Thalidomide: Consequences to Life and Economy 31
2.2.2 Structure-Activity Relationships II: Conformation. Creutzfeldt-Jakob Disease 32
2.2.2.1 Prions 32
2.2.2.2 Prions and the Creutzfeldt-Jakob Disease 33
2.2.2.3 Neurodegenerative Diseases in the Future. Consequences to Life and Economy 34
2.3 MOLECULAR MODELLING AND CONFORMATIONAL ANALYSIS 35
2.3.1 Algorithms, Storing and Analysing; "Data Mining" 36
2.3.1.1 How can I solve my problem? 36
2.3.1.1.1 The continuous problem 37
2.3.1.1.2 The discrete problems 38
2.3.1.2 How can I store my Results? 41
2.3.1.3 How can I analyze my Results? 42

3 CYCLODEXTRINS 47

3.1 INTRODUCTION 49
3.2 A SURVEY IN CYCLODEXTRINS 49
3.3 BRIEF HISTORY OF CYCLODEXTRINS 51
3.4 SYNTHESIS AND LARGE SCALE PRODUCTION 52
3.5 USES AND APPLICATIONS 54
3.6 MOLECULAR STRUCTURE. GLUCOSE: THE BASIC BRICK 58
3.6.1 Glucose 59
3.6.2 Isomers within the Aldohexose Family 59
3.6.3 Rotamers of the D-(+)-glucose 61
3.7 NOMENCLATURE AND GEOMETRICAL PARAMETERS 62
3.7.1 Lineal Chains 62
3.7.2 Cyclodextrins 62
3.7.2.1 Semisystematic Names 63
3.7.2.2 Systematic Names 64
3.7.3 Names and Geometrical Parameters 64
3.7.3.1 Simplicity 65
3.7.3.2 Compatibility 65
3.8 BENCHMARK: COMMON AND LARGE CYCLODEXTRINS 66
3.8.1 The Set of Molecules 66
3.8.2 Hydrogen Bonding and Secondary Structure 67

4 RESULTS I: METHODOLOGY. NOMENCLATURE AND DATA MINING IN CYCLODEXTRINS 71

4.1 CLASSIFICATION: INTRODUCTION 73
4.2 MOLECULAR DESCRIPTOR: NOMENCLATURE 73

4.2.1 Molecular Descriptors and 3D structure. 74
4.2.2 Descriptor I: Nomenclature by Dr. Itziar Maestre. 75
4.2.2.1 Rules of Precedence: 76
4.2.3 Descriptor II: Nomenclature in this work. 77
4.2.3.1 Rules of Precedence: 78
4.2.4 Principal Component Analysis (PCA). Checking Descriptor II. 80
4.2.4.1 "Good enough is good". 80
4.2.4.2 Principal Component Analysis. 81
4.2.4.3 Test of comparison: PCA vs. Descriptor II. 81
4.2.4.4 CD5 83
4.2.5 Estimating Total Number of Conformers: Equations. 88
4.3 CLASSIFICATION: GEOMETRIC CRITERIA 93
4.3.1 Data Mining in Conformational Pool 93
4.3.2 Conformational Search: Saturation Approach to SA and MD. 95
4.3.2.1 Deduction 95
4.3.2.2 Conclusion 96
4.3.2.3 Data Mining: Saturation Diagram 96
4.3.3 Study of Stability: Trajectories Overlapping Ratio in MD. 101
4.3.3.1 Deduction 102
4.3.3.1.1 Full Conformational Space Overlapping –near 100%-. 103
4.3.3.1.2 No Conformational Space Overlapping –0%-. 105
4.3.3.1.3 Partial Conformational Space Overlapping –Q%-. 106
4.3.3.2 Discussion and Equations. 108
4.3.3.2.1 Overlapping Equation I: Index ω 108
4.3.3.2.2 Overlapping Equation II: Index λ 110
4.4 INTEGRATED PROTOCOL FOR CONFORMATIONAL SEARCH. 111

5 RESULTS II: CONFORMATIONAL SEARCH 113

5.1 CONFORMATIONAL SEARCH: 115
5.2 SIMULATED ANNEALING (THERMAL SHOCK APPROACH) 115
5.2.1 Slow SA [300 ps]. 1024 steps 115
5.2.1.1 CD5 116
5.2.1.1.1 Saturation Analysis 116
5.2.1.1.2 Principal Component Analysis 118
5.2.1.1.3 Markov Analysis 118
5.2.1.1.4 Results 120
5.2.1.2 CD5x 120
5.2.1.2.1 Saturation Analysis 120
5.2.1.2.2 Principal Component Analysis 121
5.2.1.2.3 Markov Analysis 124
5.2.1.2.4 Results 124
5.2.1.3 CD6 (α-CD) 125
5.2.1.3.1 Saturation Analysis 125
5.2.1.3.2 Principal Component Analysis 126
5.2.1.3.3 Markov Analysis 127
5.2.1.3.4 Results 128
5.2.1.4 CD7 (β-CD) 128
5.2.1.4.1 Saturation Analysis 128
5.2.1.4.2 Principal Component Analysis 130
5.2.1.4.3 Markov Analysis 132
5.2.1.4.4 Results 133
5.2.1.5 CD8 (γ-CD) 134
5.2.1.5.1 Saturation Analysis 134
5.2.1.5.2 Principal Component Analysis 136
5.2.1.5.3 Markov Analysis 138
5.2.1.5.4 Results 140
5.2.1.6 Large Cyclodextrins: CD14, CD21, CD26 and CD28 141
5.2.2 Fast SA [30 ps] 4096 steps + MM. 142
5.2.2.1 Large Cyclodextrins: CD14, CD21, CD26 and CD28 144
5.3 GROUP PROPERTIES 145
5.3.1 Histograms of Energy 146
5.3.2 Surface Area 149
5.3.3 Radius of Gyration 151
5.3.4 Hydrogen Bonding 153

5.3.5 The rigid & flexible families *156*
5.3.5.1 Structure Superposition 156
5.3.5.2 Average structures 158
5.3.6 Saturation Approach and Equation of Conformations *160*
5.4 SUMMARY OF RESULTS 161
5.5 STRUCTURE SELECTION FOR MOLECULAR DYNAMICS 162

6 RESULTS III: MOLECULAR DYNAMICS "IN VACUO" 165

6.1 MOLECULAR DYNAMICS 167
6.2 MOLECULAR DYNAMICS "IN VACUO". 168
6.2.1 CD5 ..... *169*
6.2.1.1 Population Analysis 169
6.2.1.2 Overlapping Analysis I: Accumulative Lambda Index 170
6.2.1.3 Overlapping Analysis II: Omega Index 171
6.2.1.4 Average Structures 172
6.2.2 CD5x *174*
6.2.2.1 Population Analysis 174
6.2.2.2 Overlapping Analysis I: Accumulative Lambda Index 176
6.2.2.3 Overlapping Analysis II: Omega Index 177
6.2.2.4 Average Structures 178
6.2.3 CD6 (α-CD) *179*
6.2.3.1 Population Analysis 179
6.2.3.2 Overlapping Analysis I: Accumulative Lambda Index 180
6.2.3.3 Overlapping Analysis II: Omega Index 182
6.2.3.4 Average Structures 182
6.2.4 CD7 (β-CD) *184*
6.2.4.1 Population Analysis 184
6.2.4.2 Overlapping Analysis I: Accumulative Lambda Index 185
6.2.4.3 Overlapping Analysis II: Omega Index 187
6.2.4.4 Average Structures 187
6.2.5 CD8 (γ-CD) *188*
6.2.5.1 Population Analysis 189
6.2.5.2 Overlapping Analysis I: Accumulative Lambda Index 190
6.2.5.3 Overlapping Analysis II: Omega Index 192
6.2.5.4 Average Structures 193
6.2.6 CD14 ..... *194*
6.2.6.1 Population Analysis 195
6.2.6.2 Overlapping Analysis I: Accumulative Lambda Index 196
6.2.6.3 Overlapping Analysis II: Omega Index 198
6.2.6.4 Average Structures 198
6.2.7 CD21 ..... *200*
6.2.7.1 Population Analysis 201
6.2.7.2 Overlapping Analysis I: Accumulative Lambda Index 202
6.2.7.3 Overlapping Analysis II: Omega Index 204
6.2.7.4 Average Structures 204
6.2.8 CD26 ..... *206*
6.2.8.1 Population Analysis 206
6.2.8.2 Overlapping Analysis I: Accumulative Lambda Index 208
6.2.8.3 Overlapping Analysis II: Omega Index 210
6.2.8.4 Average Structures 210
6.2.9 CD28 ..... *211*
6.2.9.1 Population Analysis 211
6.2.9.2 Overlapping Analysis I: Accumulative Lambda Index 213
6.2.9.3 Overlapping Analysis II: Omega Index 215
6.2.9.4 Average Structures 215
6.3 GROUP PROPERTIES I: THE EFFECT OF THE SET OF CHARGES IN CD5 216
6.3.1 The AMBER philosophy: The "unit" *217*
6.3.2 Comparative Molecular Dynamics *218*
6.3.3 Results *220*
6.4 SUMMARY OF RESULTS 220

7 RESULTS IV: MOLECULAR DYNAMICS IN SOLVENT 223

7.1 MOLECULAR DYNAMICS 225
7.2 MOLECULAR DYNAMICS IN WATER SOLVENT (TIP3P H_20) 226

7.2.1 CD8 (γ-CD) *228*
7.2.1.1 Population Analysis 228
7.2.1.2 Overlapping Analysis I: Accumulative Lambda Index 230
7.2.1.3 Overlapping Analysis II: Omega Index 232
7.2.1.4 Average Structures 233
7.2.2 CD14 ..... *234*
7.2.2.1 Population Analysis 234
7.2.2.2 Overlapping Analysis I: Accumulative Lambda Index 236
7.2.2.3 Overlapping Analysis II: Omega Index 238
7.2.2.4 Average Structures 238
7.2.3 CD21 ..... *239*
7.2.3.1 Population Analysis 239
7.2.3.2 Overlapping Analysis I: Accumulative Lambda Index 241
7.2.3.3 Overlapping Analysis II: Omega Index 243
7.2.3.4 Average Structures 243
7.2.4 CD26 ..... *244*
7.2.4.1 Population Analysis 244
7.2.4.2 Overlapping Analysis I: Accumulative Lambda Index 246
7.2.4.3 Overlapping Analysis II: Omega Index 248
7.2.4.4 Average Structures 248
7.2.5 CD28 ..... *249*
7.2.5.1 Population Analysis 250
7.2.5.2 Overlapping Analysis I: Accumulative Lambda Index 252
7.2.5.3 Overlapping Analysis II: Omega Index 254
7.2.5.4 Average Structures 254
7.3 MOLECULAR DYNAMICS IN BENZENE SOLVENT (C_6H_6) 255
7.3.1 CD26 ..... *257*
7.3.1.1 Population Analysis 257
7.3.1.2 Overlapping Analysis I: Accumulative Lambda Index 259
7.3.1.3 Overlapping Analysis II: Omega Index 262
7.3.1.4 Average Structures 262
7.4 GROUP PROPERTIES II: THE EFFECT OF SOLVENT IN CD26 263
7.4.1 Brief summary of solvent-solute interactions *263*
7.4.2 Comparative Molecular Dynamics *264*
7.4.2.1 Total Number of Conformations according to D-II 264
7.4.2.2 Solvation Energy 268
7.4.2.3 Hydrogen Bonds Occupancy Ratio 268
7.4.2.4 Average Structures 270
7.4.3 Results *271*
7.5 SUMMARY OF RESULTS 272

8 CONCLUSIONS 275

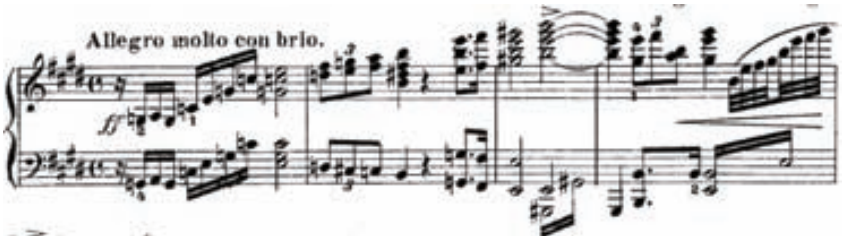
8.1 QUESTION I 277
8.1.1 Conclusion I *277*
8.2 QUESTION II 277
8.2.1 Conclusion II *277*

9 APPENDIX: METHODOLOGY 281

9.1 INTRODUCTION 283
9.2 MOLECULAR MODELLING 283
9.3 MOLECULAR MECHANICS 284
9.3.1 Classical Equations *284*
9.3.2 Force Fields *286*
9.3.2.1 List of Force Fields 286
9.3.2.2 Parametrisation 288
9.3.3 Solvation Models *289*
9.3.3.1 Implicit Solvation Models 290
9.3.3.2 Explicit Solvation Models 292
9.3.3.3 Periodic Boundary Conditions (PBC) 294
9.3.3.4 Mixed Solvation Models: MM-PBSA 295
9.3.3.5 Parameterisation 296
9.4 METHODOLOGIES 297
9.4.1 Geometrical Optimization *297*
9.4.2 Global Search Methods and Pool of Conformations *299*

9.4.2.1 The Scale-factor in conformational problems 299
9.4.2.1.1 Topology 300
9.4.2.1.2 Folding and Aggregation 300
9.4.2.2 Genetic Algorithms 302
9.4.2.2.1 Holland's GA Schema Theorem 302
9.4.2.2.2 GACK and Macromodel 308
9.4.2.3 Simulated Annealing and Monte Carlo 310
9.4.2.3.1 Classical Algorithm 311
9.4.2.3.2 Equivalent Thermal Algorithm 312
9.4.2.3.3 Markov Chains associated to GA, SA and MC 315
9.4.2.4 Molecular Dynamics 316
9.4.2.4.1 Theoretical Background 317
9.4.2.4.2 Algorithmic Implementation 319
9.4.2.4.3 Sampling and Conformational Information 320

OBJECTIVES



Richard Strauss
Don Juan op. 20
Tone Poem for Large Orchestra
(1889)

"Nothing great was ever achieved without enthusiasm."

Ralph Waldo Emerson

1 OBJECTIVES

The research work presented in this memory was initially planned as the final part of a wide range study on cyclodextrins (CDs) developed by former members of our group. Initially carried out by Dr. Ivan Beà and continued by Dr. Itziar Maestre, this work meant the culmination of that project.

Results obtained during Dr. Beà's research period were submitted for publication at that time, but referees considered that some of them should be revised in order to ensure the full validity of the conclusions: We were sent a report comprising the two principal objections about the work. Briefly summarizing its contents, the reasons exposed to support the non-acceptance of the paper are shown here:

1) The behaviour of the system –large cyclodextrins– highly depends on the Force Field parameters employed in the calculations. Therefore, it should be proved that results in the paper are in accordance to those obtained working with different Force Fields to ensure full generality.

2) The data related to Molecular Dynamics (MD) calculations seemed to show a tight dependence on starting conformations and, apparently, it leads to the conclusion that the explorations of the conformational space of the systems under study were inefficiently done. This –meaning also a lack of generality- should be revised in order to assure the validity of the conclusions.

Therefore, these questions had to be extensively revised to fulfil the integrity of the work. The first point was studied in depth by Dr. Itziar Maestre in her Thesis work[3], concluding that *"The differences were not significant when parm99 and glycam2000a were used."* And also *"Parm94 and MM3* results differ considerably due to inadequate bending parameters (parm94) or to a strong stabilization by the intramolecular hydrogen bonds between hydroxyl groups."* [4]

The second point still remained unanswered in those days and soon became the starting point of the present research. Day after day, as we were involved in the development of

[3] Maestre, I.; *Doctoral Thesis*. Chemistry Department. UAB. **2004**.
[4] Maestre, I.; Beà, I.; Ivanov, P.M.; Jaime, C.; *Theor Chem Acc*. **2007**. *117(1)*. 85-97.

the study, trying to guess the best way to address the problem with MD, we realised that the key point was the Conformational Search.

Folding is extremely important in large molecules because their function and properties are closely related to it. Nevertheless, the total number of conformations is exceedingly high to ensure optimal prediction. This assumption was growing in importance and, finally, it forced us to change the point of view of the whole work, focusing our interests in developing a new conformational search methodology rather than merely study a set of macromolecules.

Conformational studies are especially relevant in several areas of Science other than Chemistry: mainly Biochemistry, Medicine and Biology (i.e. protein activity –Enzymes- is tightly related to 3D structure as will be shown in the next chapter).

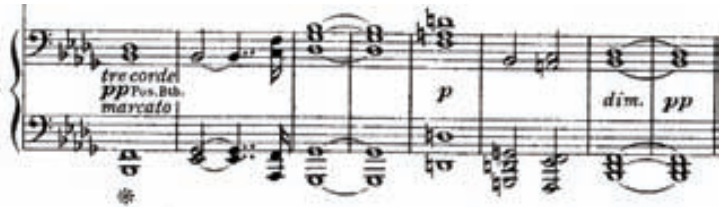
There are enough reasons, therefore, to frame our objectives into this context.

Summarizing, the purpose of this Thesis is the quest for the best answers to the following statements:

I) The initial problem regarding Molecular Dynamics inefficient exploration of the conformational space suggested us the necessity of developing a general methodology in the area of knowledge of "Conformational Search and MD analysis" as the first objective of this work.

II) As we use a full family of cyclodextrins (CDs) as a benchmark to evaluate the efficiency of our methodology, the choice of these molecules will also allow us to give a proper answer to the "second objection" to Dr. Beà's work, being this the second general objective of the present Thesis.

INTRODUCTION



Richard Strauss
An Alpine Symphony op. 64
Tone Poem for Large Orchestra
(1915)
Night (The Mountain's Leitmotiv)

"There is a long journey ahead..."

2 INTRODUCTION: CONFORMATIONAL IMPORTANCE IN CHEMISTRY

2.1 Chemistry beyond the Reactions

2.1.1 "In the beginning God created Chirality…"

New discoveries on abiogenesis -the origin of life on Earth- shake public opinion from time to time. Although several plausible theories are currently accepted there is a question that still remains unclear: the fact of Homochirality and life[5].

In biology, homochirality is found inside living organisms. I.e. most active forms of amino acids are of the L-form (D-serine being a notable exception) and many biologically relevant sugars are of the D-form[6] (being L-arabinose also another exception). Moreover, it is known that the alternative form of chiral bioactive molecules is inactive and sometimes even toxic.

The origin of this phenomenon is not clearly understood and it is even unclear whether homochirality has a purpose. Anyway, we are not concerned about it; we are focusing our interest in its direct consequences to life. The only truth is that *"asymmetry"* is a constant in Universe, and also in physics, chemistry and biochemistry. This means that, beginning from subatomic particles[7], going through the smaller molecules up to the larger compounds, stereospatial properties will drive the behaviour of every substance, irrespective of its nature.

Starting from the origins of chirality, going through modern stereochemistry until we reach proteins, foldamers, and molecular recognition, a few examples of medical interest will be presented in order to frame the guideline of this work: the importance of spatial chemistry beyond chirality, and hence clear the path to the proposal of a new methodological approach to conformational analysis, in Chapter IV.

---

[5] (a) Bonner, W.A.; *Origins of Molecular Chirality*. In *Exobiology*. (Ed. Ponnamperuma, C.) **1972**. ISBN 0444101101. (b) Miller, S.L.; Orgel, L.E.; *The Origins of Life on the Earth*. Prentice Hall. **1974.** ISBN-13: 978-0136420743. (c) Ulbricht, T.L.V.; *Origins of Life*. **1981**. *11(1-2)*. 55-70. (d) Mathew, S.P.; Iwamura, H.; Blackmond, D.G.; Angewandte Chemie. **2004**. *116(25)*. 3379-3383.

[6] (a) Lehninger, A.L.; Nelson, D.L.; Cox, M.M.; *Principles of Biochemistry*. 4th Ed. W.H. Freeman & Company. **2004**. ISBN 9780716743392. (b) Mason, S.F.; *Nature*. **1984**. *311(5981)*. 19-23.

[7] (a) Bromley, D.A.; *Gauge Theory of Weak Interactions*. Springer. **2000**. ISBN 3-540-67672-4. (b) Kane, G.L.; *Modern Elementary Particle Physics*. Perseus Books. **1987**. ISBN 0-201-11749-5.

2.1.2 Brief History of Chirality: "*The Birth of Stereochemistry"*

It was difficult to guess the impact that stereochemistry[8] would have in the forthcoming years when the mineralogist René Just Haüy (1743–1822) opened the door to *"spatial chemistry"* discovering in 1801 the enantiomorphic quartz crystals[9]: nonsuperimposable complete mirror images of each other [Fig. 2.1 (a)].

On the other hand, in 1809, engineer-physicist-mathematician Etienne-Louis Malus (1775–1812) discovered the polarisation of light by reflection [Fig. 2.1 (b)]. And later, Sir David Brewster (1781–1868) quantified this effect –Brewster's Law[10]- publishing his results in 1812.

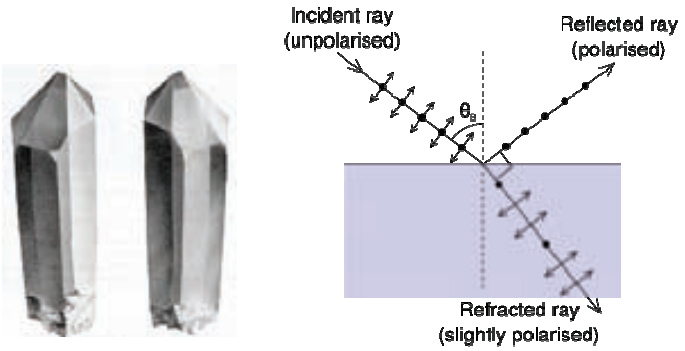


Figure 2.1: (a) Left and right quartz crystals. (b) Polarisation by reflection.

In the following years, three scientists made significant advances in a short period of time when they separately found a relationship between these two phenomena: In 1811, François Jean Dominique Arago (1786–1853) observed the optical rotation in quartz. Later in 1819, Eilhard Mitscherlich (1794–1863) enunciated his law of isomorphism, which states that compounds crystallizing together probably have similar structures and compositions. And Jean-Baptiste Biot (1774–1862) found in 1815 optical rotation in some crystals, but also in liquids and solutions (tartaric acid, glucose and camphor). The latter was a crucial finding because it revealed that this property was not exclusively

[8] Quiroga Feijóo, M.L.; *Estereoquímica. Conceptos y Aplicaciones en Química Orgánica*. Editorial Síntesis. **2007**. pg: 12-13. ISBN 978-84-975650-6-6.
[9] Haüy, R.J. *Traité de Minéralogie* (5 vols). **1801**. Bibliothèque Nationale de France, Gallica.
[10] Sears, F.W.; Zemansky, M.W.; Young, H.D.; *Física Universitaria*. Addison-Wesley Iberoamericana, 6ª Ed. en Español. **1986**. ISBN 0-201-64013-9.

associated with the architecture of the crystals or the states of aggregation, thus paving the way for an underlying molecular explanation.

Finally, a few decades later, scientist Louis Pasteur (1822–1895) made also a significant discovery when he resolved in 1847 the problem concerning the nature of tartaric acid, establishing the molecular foundations of this macroscopic phenomenon[11]. Upon detailed examination of the minute crystals of sodium ammonium tartrate, Pasteur noticed that they came in two asymmetric forms that were mirror images of one another and correctly deduced that the asymmetry in crystals –macroscopic viewpoint- was an evidence of the asymmetry in the molecules –microscopic viewpoint-. Hence, molecules could exist in two different, *"dissymmetric"* or *"enantiomeric"* forms [Fig. 2.2].

Figure 2.2: Crystals of sodium ammonium tartrate.

Anyway, not a single hint on the clue to molecular bonding or atomic spatial arrangement had been given yet. The answer to this question was proposed independently and almost at the same time in 1874 by Jacobus Henricus van't Hoff (1852–1911) and Joseph Achille Le Bel (1847–1930). Both of them considered the tetrahedrally-bonded carbon atoms as the origin of the optical activity when all four substituents were different [Fig. 2.3].

[11] Pasteur, L. **Doctoral Thesis.** *Pasteur Œuvre tome 1 – Dissymétrie Moléculaire*. **1847**. Bibliothèque Nationale de France, Gallica.

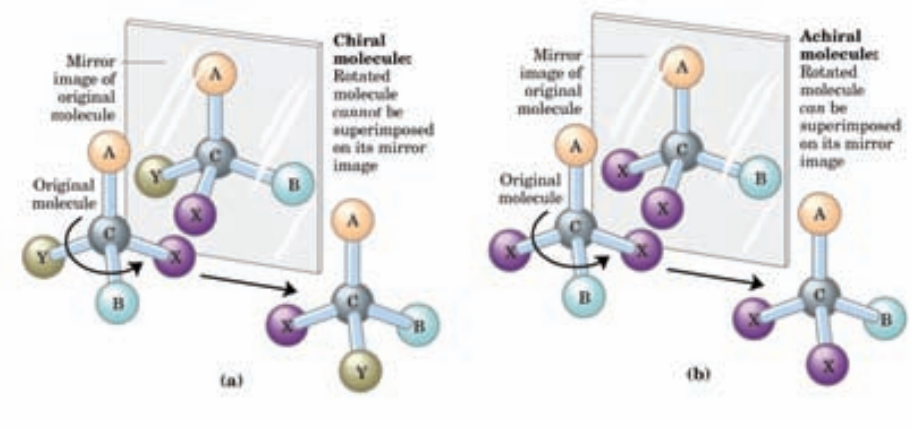


Figure 2.3: *"Chiral"* and *"Achiral"* molecules[6].

The next contribution to this field was made in 1893 by Lord Kelvin (1824–1907) who proposed the term *"chiral"* (from Greek word for hand: *"χειρ"*) to denote the *"dissymmetry"* in the tetrahedral carbon atom C*abcd [Fig. 2.4].

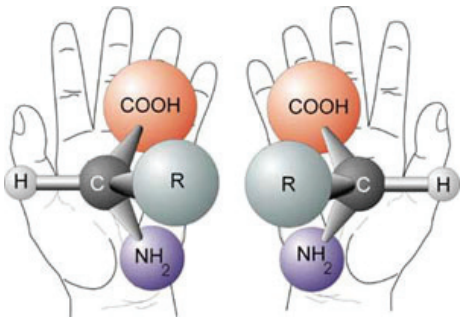


Figure 2.4: Etymology, *"Chiral"* molecules, from Greek *"χειρ"*.

So far, we have seen the relationships between optical rotation and *chiral* carbon atoms, but soon were unveiled other types of molecules, like phosphines and sulphoxides, which showed this behaviour being free of them [Fig. 2.5]. Thus, the definition of chirality had to be revised and updated again to extend its range of applicability.

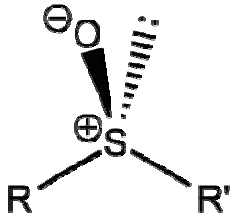


Figure 2.5: Chirality in absence of tetrahedral carbon atoms.

In general, a molecule is *achiral* (not *chiral*) if and only if it has an axis of improper rotation; that is, an n-fold rotation (rotation by 360°/n) followed by a reflection in the plane perpendicular to this axis that maps the molecule onto itself. Further contributions were made by Cahn, Ingold and Prelog when they established in 1966 the conventional sequence rules[12] for designating unambiguously the priority of the ligands. Later in 1984, Mislow and Siegel[13] introduced the general concepts *stereocenter* and *stereogenic centre* meaning any atom in a molecule bearing groups such that an interchanging of any two of them led to a stereoisomer. This definition allowed an extended vision in which atoms other than carbon might also be considered as generators of chirality.

2.1.3 Stereochemistry today

Stereochemistry is more than just chirality[14]. IUPAC comprises several types of *isomers*[15], including *constitutional isomers*, *stereoisomers*, *enantiomers*, *diastereomers*, *cis-trans isomers*, *conformers* and *rotamers* [Fig. 2.6].

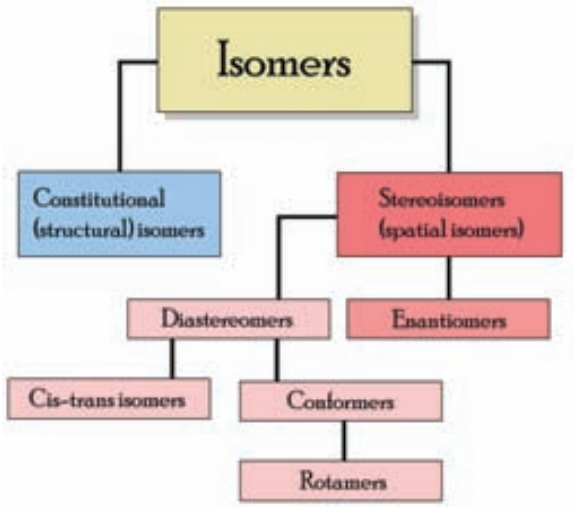


Figure 2.6: Chirality Flow Chart.

[12] (a) Cahn, R.S.; Ingold, C.K.; Prelog, V.; *Angew. Chem. Int. Ed. Eng*. **1966**. *5(4)*. 385-415, 511 (b) Prelog, V.; Helmchen, G.; *Angew. Chem. Int. Ed. Eng*. **1982**. *21(8)*. 567-583.
[13] Mislow, K.; and Siegel J.; *J. Am. Chem. Soc.* **1984**. *106(11)*. 3319-3328.
[14] Testa, B.; *Principles of Organic Stereochemistry*. Marcel Dekker, INC. **1979**. ISBN 0-8247-6758-6
[15] (a) IUPAC; *Pure & Appl. Chem.*. **1976**. *45(1)*. 11-30. (b) IUPAC; *Pure & Appl. Chem.*. **1996**. *68(12)*. 2193-2222. (c) Gold, V.; Loening, K.L.; McNaught, A.D. and Shemi, P. *The Gold Book: Compendium of Chemical Terminology*. Blackwell Science, **1987**. ISBN 0-63201-7651(8). (d) McNaught, A.D. and Wilkinson, A. *The Gold Book: Compendium of Chemical Terminology, 2nd edition*. Blackwell Science, **1997**. ISBN 0-86542-6848. (e) http://old.iupac.org/publications/compendium/index.html. *"The Gold Book"* on-line.

Nevertheless, in recent years a new group of molecules have appeared exhibiting optical behaviour on the basis of exotic molecular stereospatial properties such as *axial chirality* (*allenes*, *cumulenes*, *spiranes*, *atropisomers* like *biphenyls* and *binols*…), *planar chirality* (*cyclophanes*…), and *helicoidal chirality* (*helicenes*).

In recent years, *mechanical bondings* have led stereochemistry to a New Universe of *molecular topologies*[16] (*rotaxanes*[17], *catenanes*[18], *knotanes*[19]…) that in most cases exceed the limits of our imagination (See examples ahead: [Fig. 2.13] and [Fig. 2.14]).

2.1.4 Supramolecular Chemistry: "Architecture in the Microscopic World"

Beyond the aforesaid stereochemistry, there is still a New World of large molecules showing interesting 3D properties. Supramolecular Chemistry[20] refers to the area of chemistry that focuses on the noncovalent bonding interactions of molecules.

Donald D. Cram, Charles J. Pedersen, and Jean-Marie Lehn were jointly awarded the Nobel Prize in Chemistry in 1987 for their works on Supramolecular Chemistry. The statement *"chemistry beyond the molecule"* coined by Lehn[21] is the *leitmotiv* that defines their new discoveries.

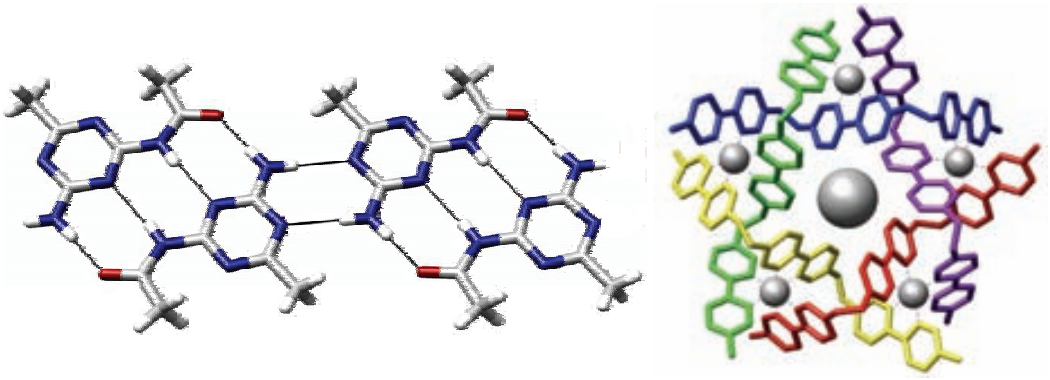
The new area of knowledge comprises several branches:

2.1.4.1 Molecular Self-Assembly

Is the assembly of molecules without guidance, management nor interference from an outside source. There are two types, namely intramolecular and intermolecular self-

[16] Breault, G.A.; Hunter, C.A.; Mayers, P.C.; *Tetrahedron*. **1999**. *55(17)*. 5265-5293.
[17] Stoddart, J.F.; *Nature*. **1988**. *334(6177)*. 10-11.
[18] Safarowsky, O.; Windisch, B.; Mohry, A.; Vögtle. F.; *Journal für praktische Chemie*. **2000**. *342(5)*. 437-444.
[19] Lukin, O.; Vogtle, F.; *Angew. Chem. Int. Ed. Eng*. **2005**. *44 (10)*. 1456-1477.
[20] (a) Lehn, J.M.; *Science*. **1993**. *260(5115)*. 1762-1763. (b) Steed, J.W.; Atwood, J.L.; *Supramolecular Chemistry*. John Wiley & Sons. **2000**. ISBN 0-471-98791-3.
[21] Lehn, J.M.; *Supramolecular Chemistry - Concepts and Perspectives*. John Wiley & Sons. **1995**. ISBN 3-527-29311-6.

assembly. Most often the term molecular self-assembly refers to intermolecular self-assembly (examples from Beijer and Lehn) [Fig. 2.7], while the intramolecular analogue is more commonly called folding.



(a) **(b)**

Figure 2.7: Examples of Molecular Self-Assembly by (a) Beijer[22] and (b) Lehn[23].

2.1.4.2 Folding: Foldamers and Proteins

2.1.4.2.1 Foldamers

A foldamer is a discrete chain molecule or oligomer that adopts a secondary structure [Fig. 2.8] stabilized by non-covalent interactions[24]. They are artificial molecules that mimic the ability of proteins, nucleic acids, and polysaccharides to fold into well-defined conformations, such as helices and β-sheets.

Foldamers have been shown to display a number of interesting supramolecular properties including molecular self-assembly, molecular recognition, and host-guest chemistry. They are studied as models of biological molecules and they also have great potential application to the development of new functional materials.

[22] Beijer, F.H.; Kooijman, H.; Spek, A.L.; Sijbesma, R.P.; Meijer, E.W.; *Angew. Chem. Int. Ed. Eng*. **1998**. *37(1-2)*. 75-78.
[23] Hasenknopf, B.; Lehn, J.M.; Kneisel, B.O.; Baum, G.; Fenske, D.; *Angew. Chem. Int. Ed. Eng*. **1996**. *35(16)*. 1838-1840.
[24] (a) Gellman, S.H.; *Acc. Chem. Res*. **1998**. *31(4)*. 173-180. (b) Hill, D.J.; Mio, M.J.; Prince, R.B.; Hughes, T.S.; Moore, J.S.; *Chem. Rev*. **2001**. *101(12)*. 3893-4012.

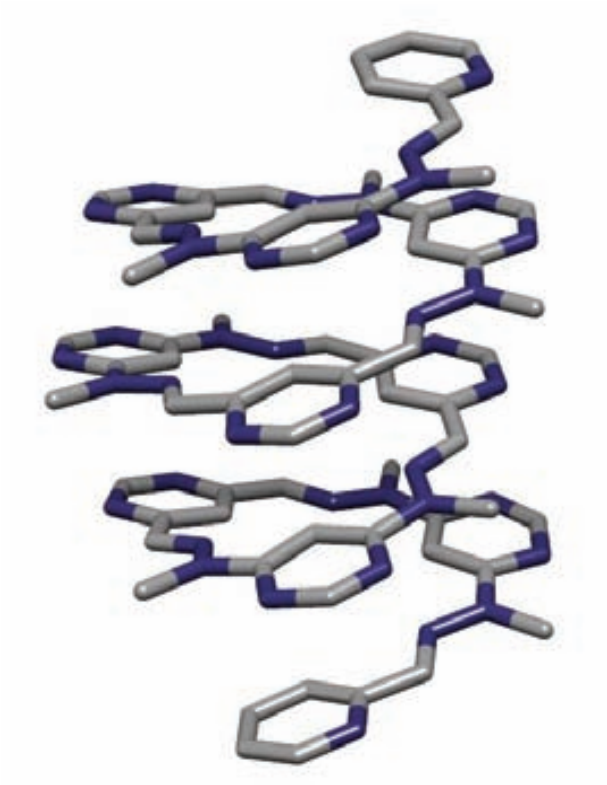


Figure 2.8: Foldamer by Lehn[25] exhibiting Secondary Structure.

2.1.4.2.2 Proteins

Proteins are linear polymers built from 20 different L-α-amino acids. They adopt an entangled complex three dimensional conformation including secondary, tertiary and quaternary structure -levels of folding- that is essential for developing their function [Fig. 2.9].

[25] Schmitt, J.L.; Stadler, A.M.; Kyritsakas, N.; Lehn, J.M.; *Helv. Chim. Acta*. **2003**. *86(5)*. 1598-1624.

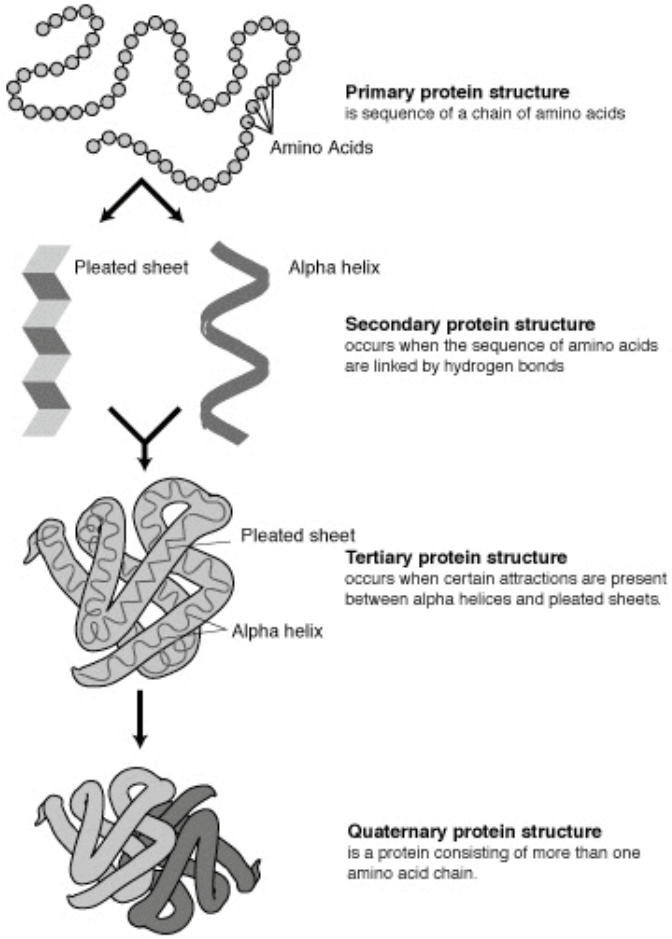


Figure 2.9: Proteins, primary structure, and folding levels.

2.1.4.2.3 Folding

Understanding folding is extremely important to know the mechanism of enzymes and drugs *in vivo*. The amino-acid sequence -or primary structure- of a protein predisposes it towards its native conformation or conformations, which is quite in agreement with ***Anfinsen's Dogma***[26]. On the other hand, ***Levinthal's Paradox***[27] states that, because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an astronomical number of possible conformations [Fig. 2.10], and thus, folding cannot be a random process. The folding mechanism also depends on the

[26] (a) Anfinsen, C.B.; *Science*. **1973**. *181(4096)*. 223-230. (b) Anfinsen C.B.; *Nobel Prize Lecture*. **1971**.
[27] Levinthal C.; *Journal de Chimie Physique et de Physico-Chimie Biologique*. **1968**. *65(1)*. 44-45.

characteristics of the cytosol, including the nature of the primary solvent (water or lipid), the concentration of salts, the temperature, and molecular chaperones.

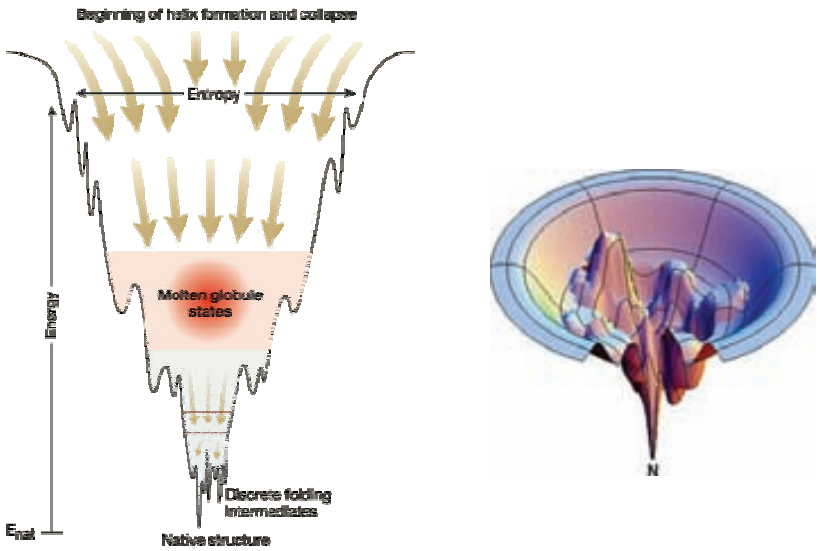


Figure 2.10: Free Energy Landscape and folding[28]. Horizontal axis is related to number of conformations and therefore Entropy, while vertical axis represents the Free Energy.

Levinthal's Paradox is also a starting point to approach the computational problem of conformational search in large macromolecules. As said, proteins do not fold via an exhaustive random search of their conformational space. Furthermore, it is also known that proteins fold spontaneously on short timescales. Therefore, an intensive, purely random search cannot succeed, as Levinthal himself believed. Heuristic and intelligent computational algorithms seem to be the best choices instead of purely random ones.

Folding and Computer Sciences are nowadays so important in Life Sciences that Stanford University Medical Centre at Stanford University is leading in coordination with the National Science Foundation and the National Institutes of Health, the project ***Folding@home***[29] jointly together with a pioneer group of computer-science companies including Google©, Dell©, Apple©, Intel©, ATI© and Simbios©.

Computational studies in conformational search are extremely time consuming and they are usually carried out in collaboration with supercomputing centres according to their highly demanding requirements. *Folding@home* is a new option; a distributed

[28] Image source: Lehninger, A.L.; Nelson, D.L.; Cox, M.M.; *Principles of Biochemistry*. 4th Ed. W.H. Freeman & Company. **2004**. ISBN 9780716743392.
[29] (a) http://www.stanford.edu/group/pandegroup/folding/about.html (b) http://folding.stanford.edu/

computing project where people from throughout the world download and run software to band together to make one of the largest supercomputers in the world. *Folding@home* uses novel computational methods coupled to distributed-computing, to simulate problems millions of times more challenging than those previously achieved. Similar technologies have been already employed by NASA and Berkeley University on SETI research also under the project name ***SETI@home***[30].

2.1.4.3 Molecular Recognition

Both, molecular recognition and host-guest chemistry are different concepts but tightly related.

The term molecular recognition, which can be divided into static and dynamic recognition, is more general and refers to the specific interaction between two or more molecules through noncovalent bonding forces such as hydrogen bonding, metal coordination, hydrophobic forces, van der Waals forces, π-π interactions, and/or electrostatic effects [Fig. 2.11].

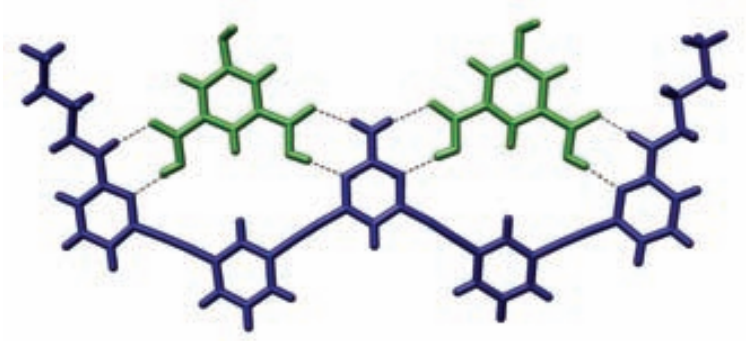
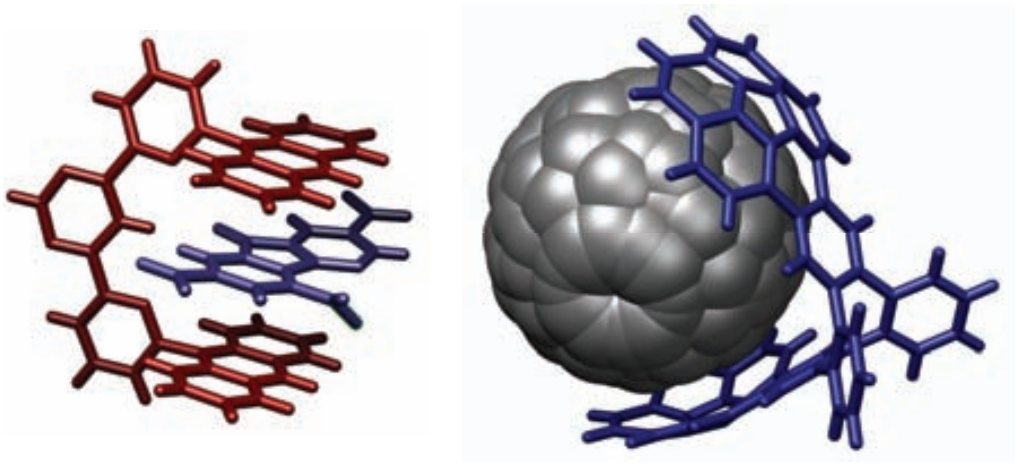


Figure 2.11: Example of Molecular Recognition by Moore[31].

Molecular tweezers, sometimes also termed molecular clips, are a classical example [Fig. 2.12 (a)]. Nevertheless, the structure recently proposed by Sygula [Fig. 2.12 (b)] is a surprisingly new supramolecular assembly that connects molecular recognition with nanomaterials represented by the fullerene.

[30] http://setiathome.berkeley.edu/
[31] Bielawski, C.; Chen, Y.S.; Zhang, P.; Prest, P.J.; Moore, J.S.; *Chem. Commun*. **1998**. *(12)*. 1313-1314.



(a) (b)

Figure 2.12: Example of Molecular Recognition by (a) Lehn[32] and (b) Sygula[33].

2.1.4.4 Host-Guest Chemistry

Host-guest chemistry describes a more specific situation where complexes are composed of two or more molecules or ions held together in unique structural relationships by hydrogen bonding, or ion pairing, or van der Waals forces other than those of full covalent bonds.

Host and guest species involved in this phenomenon exhibit molecular complementarity[34], which means that, whether statically or dynamically, at any time during the complexation process, the host must adopt a conformation that optimally enclose the guest, and vice-versa, adapting to one another. Within the complex, the host component is defined as an organic molecule or ion whose binding sites converge in the complex and the guest component is defined as any molecule or ion whose binding sites diverge in the complex.

[32] Petitjean, A.; Khoury, R.G.; Kyritsakas, N.; Lehn, J.M.; *J. Am. Chem. Soc*. **2004**. *126(21)*. 6637-6647.
[33] (a) Sygula, A.; Fronczek, F.R.; Sygula, R.; Rabideau P.W.; Olmstead, M.M.; *J. Am. Chem. Soc*. **2007**. *129(13)*. 3842-3843. (b) Wong, B.M.; *J. Comput. Chem*. **2009**. *30(1)*. 51-56.
[34] Gellman, S.H.; *Chem. Rev*. **1997**. *97(5)*. 1231-1232.

There are several examples falling into this category, i.e. *cryptands*[35], *carcerands*[36], *hemicarcerands*[37] and *carceplexes*, *calixarenes*[38], *cucurbiturils*[39] and *crown ethers*[40] [Fig. 2.13].

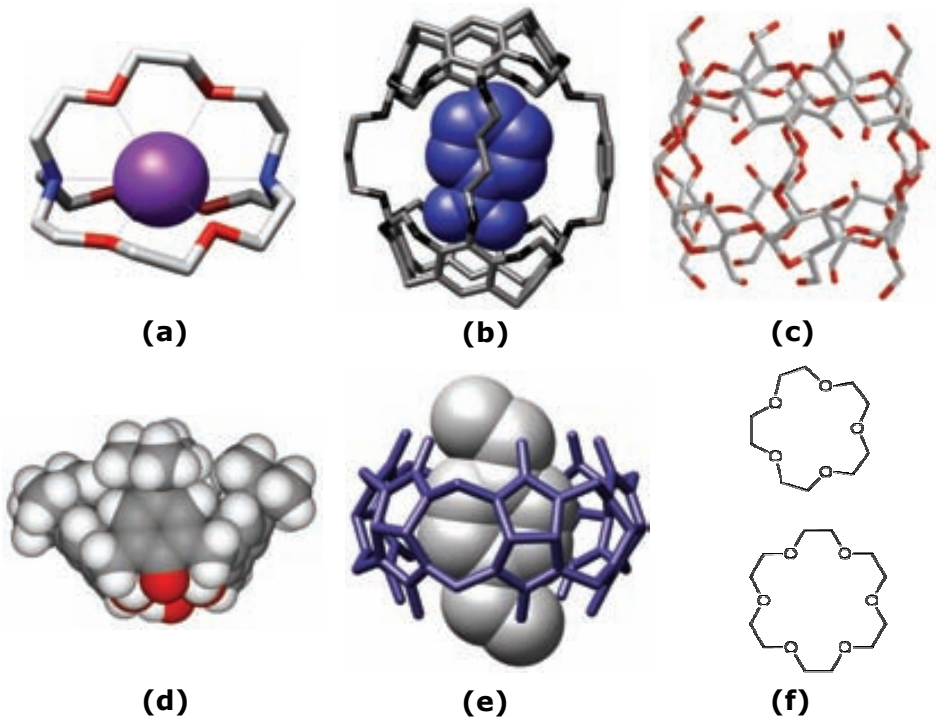


Figure 2.13: Collected examples of Host-Guest Chemistry: (a) Cryptand , (b) Hemicarcerand, (c) Carcerand, (d) Calixarene, (e) Cucurbituril and (f) Crown Ethers.

2.1.4.5 Mechanically-Interlocked Molecular Architectures

Mechanically-interlocked molecular architectures are connections of molecules not through traditional bonds, but instead as an indirect consequence of their topology. This molecular linking is analogous to keys on a key-ring. The keys are not directly connected to the key-ring but they cannot be separated without breaking it. Likewise,

[35] Alberto, R.; Ortner, K.; Wheatley, N.; Schibli, R.; Schubiger, A.P.; *J. Am. Chem. Soc*. **2001**. *123(13)*. 3135-3136.
[36] Burusco, K.K.; Ivanov, P.M.; Jaime, C.; *ARKIVOC*. **2005**. *(9)*. 287-304.
[37] Yoon, J.; Knobler, C.B.; Maverick, E.F.; Cram, D.J.; *Chem. Commun*. **1997**. *(14)*. 1303-1304.
[38] Gutsche, D.C.; *Calixarenes*. Cambridge: Royal Society of Chemistry. **1989**. ISBN 0-85186-385-X.
[39] Freeman, W.A.; *Acta. Crystallogr. B*. **1984**. *40(4)*. 382-387.
[40] Pedersen, C.J.; *J. Am. Chem. Soc*. **1967**. *89(26)*. 7017-7036.

interlocked molecules can be separated only with significant distortion of the covalent bonds that make up the conjoined supra-structure.

Examples of mechanically-interlocked molecular architectures [Fig. 2.14] include those widely studied by Stoddart[41]: *catenanes*[42], *rotaxanes*[43], *suitanes*[44], *molecular knots*[45], and *molecular Borromean rings*[46].

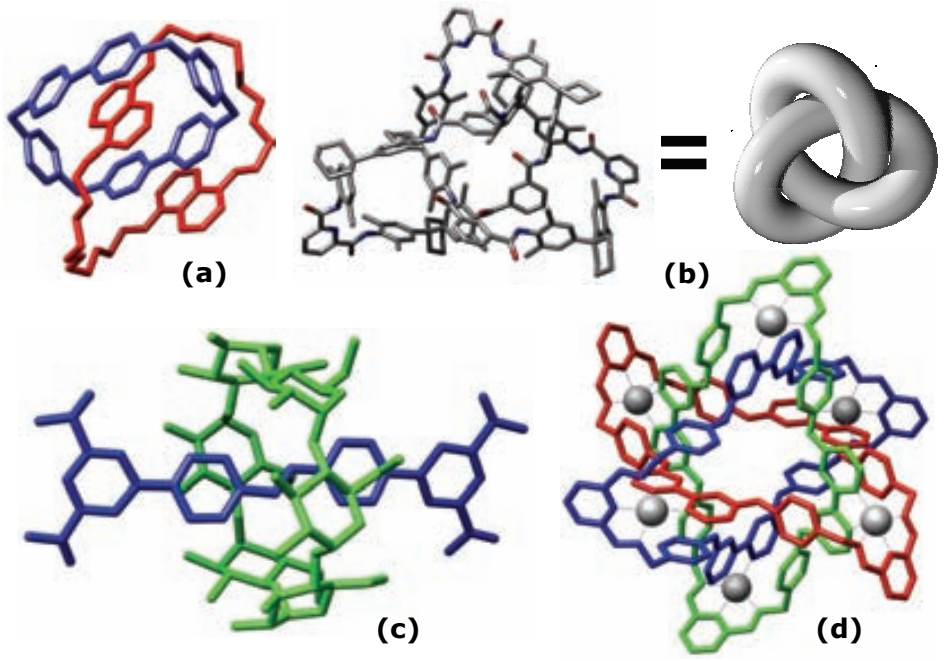


Figure 2.14: Collected examples of Mechanically-Interlocked Molecular Architectures: (a) Catenane (Stoddart), (b) Knotane, a chiral molecule (Vogtle), (c) Rotaxane (Stanier), and (d) Borromean Ring (Stoddart).

[41] J. Fraser Stoddart (b 1942): Full Work. http://stoddart.chem.ucla.edu/
[42] (a) De Federico, M.A.; *Doctoral Thesis*. Chemistry Department. UAB. **2006**. (b) Ashton, P.R.; Brown, C.L.; Chrystal, E.J.T.; Goodnow, T.T.; Kaifer, A.E.; Parry, K.P.; Philp, D.; Slawin, A.M.Z.; Spencer, N.; Stoddart, J.F.; Williams, D.J.; *J. Chem. Soc., Chem. Commun*. **1991**. *(9)*. 634-639.
[43] (a) Grabuleda, X.; *Doctoral Thesis*. Chemistry Department. UAB. **2000**. (b) De Federico, M.A.; *Doctoral Thesis*. Chemistry Department. UAB. **2006**. (c) Pérez, J.; *Doctoral Thesis*. Chemistry Department. UAB. **2008.** (c) Stanier, C.A.; Connell, M.J.O.; Anderson, H.L.; Clegg, W.; *Chem. Commun*. **2001**. *(5)*. 493-494.
[44] Williams, A.R.; Northrop, B.H.; Chang, T.; Stoddart, J.F.; White, A.J.P.; Williams, D.J.; *Angew. Chem. Int. Ed. En*. **2006**. *45(40)*. 6665-6669.
[45] (a) Albrecht-Gary, A.M.; Dietrich-Buchecker, C.O.; Guilhem, J.; Meyer, M.; Pascard, C.; Sauvage, J.P.; *Recl. Trav. Chim. Pays-Bas*. **1993**. *112(6)*. 427-428. (b) Safarowsky, O.; Nieger, M.; Frohlich, R.; Vogtle, F.; *Angew. Chem. Int. Ed. En*. **2000**. *39(9)*. 1616-1618.
[46] Chichak, K.S.; Cantrill, S.J.; Pease, A.R.; Chiu, S.H.; Cave, G.W.V.; Atwood, J.L.; Stoddart, J.L.; *Science*. **2004**. *304(5675)*. 1308-1312.

2.2 Stereochemistry: Importance of the molecular spatial arrangement

A wide variety of molecules exhibiting different stereospatial arrangements has been shown. In some cases those species were rarities whose applicability and interest are hardly appreciated beyond universities. Things changed when some crucial events regarding medical interest were made public in the 60's and 90's of the 20th century.

At this point, it is not difficult to find a pair of examples that clearly show the unquestionable importance of stereochemistry in human life and in the economy: On the one hand, ***chirality***, the ordinary one associated to asymmetric carbon atoms, is unfortunately the most well known example due to *"The Thalidomide Disaster"*. On the other hand, there is a no less important example in recent years in the field of ***molecular conformations***: the *"Creutzfeldt-Jakob Disease"*.

2.2.1 Structure-Activity Relationships I: Chirality

Pharmaceutical activity *in vivo* is closely related to spatial arrangement of atoms and functional groups due to molecular recognition[47]. Enzymes and proteins in general possess specifically designed *"pockets"* to bind certain molecules with an extremely high level of stereoselectivity [Fig. 2.15]. Hermann Emil Fischer (1852–1919) proposed in 1894 that the reactions of chiral substances are governed by a *"key and lock"* principle, the particular product resulting from the best stereochemical fit being naturally selected. This behaviour is known as *chiral biodiscrimination*.

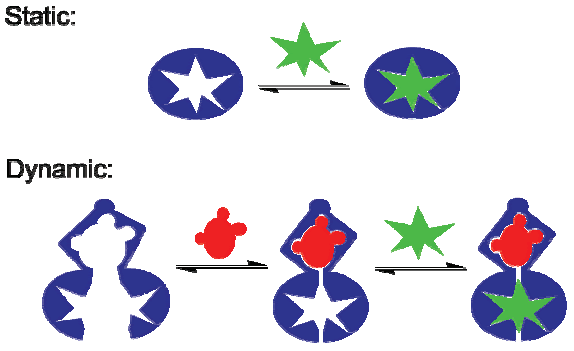


Figure 2.15: Example of molecular recognition: static and dynamic.

[47] Gellman, S.H.; *Chem. Rev*. **1997**. *97(5)*. 1231-1232.

Hence, activity is not necessarily the same for each enantiomer in a pair and there are sometimes notorious differences in behaviour[48]. When this occurs, the most active of them is called *eutomer*, whilst the less active is the *distomer*. According to this, there are four possibilities considering the differential activity ratio[49].

2.2.1.1 Specific biological activity for a single enantiomer

For example, levodopa[50] is the active enantiomer used as a prodrug to increase dopamine levels for the treatment of Parkinson's disease, since it is able to cross the blood-brain barrier, whereas dopamine itself cannot. Once levodopa has entered the central nervous system (CNS), it is metabolized to dopamine [Fig. 2.16].

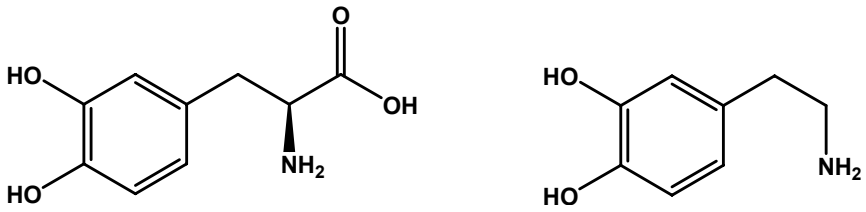


Figure 2.16: Levodopa (left) and Dopamine (right).

2.2.1.2 Identical qualitative and quantitative activity for any enantiomer

This indirect effect is sometimes present if enantiomers can interconvert *in vivo*. For example, if a human is given pure the *distomer* it converts into the *eutomer*. It is the case of ibuprofen[51], a non-steroidal anti-inflammatory drug [Fig. 2.17]. In fact, it was found that (*S*)-(+)-ibuprofen (dexibuprofen) was the active form both *in vitro* and *in*

[48] (a) Nguyena, L.A.; Heb, H.; Pham-Huyc, C.*; *Int. J. Biomed. Sci*. **2006**. *2(2)*. 85-100. (b) Slováková, A.; Hutt, A.J.; *Ceska Slov Farm*. **1999**. *48(3)*. 107-112. (c) Ehrlich, G.E.; *Am J Hosp Pharm*. **1992**. *49(9 Suppl 1)*. 15-18. (d) Sinko, G.; *Arh Hig Rada Toksikol*. **2005**. *56(4)*. 351-361. (e) Brocas, D.R.; Jamali, F.; *Pharmacotherapy*. **1995**. *15(5)*. 551-564. (f) Islam, M.R.; Mahdi, J.G.; Bowen, I.D.; *Drug Saf*. **1997**. *17(3)*. 149-165.

[49] Quiroga Feijóo, M.L.; *Estereoquímica. Conceptos y Aplicaciones en Química Orgánica*. Editorial Síntesis. **2007**. pg: 118-120. ISBN 978-84-975650-6-6.

[50] (a) **IUPAC**: (*S*)-2-amino-3-(3,4-dihydroxyphenyl)propanoic acid. (b) Carlsson, A.; Lindqvist, M.; Magnusson T.; *Nature*. **1957**. *180(4596)*. 1200-1200. (c) Abbott, A.; *Nature*. **2007**. *447(7143)*. 368-370. (d) Benes, F.M.; *Trends in Pharmacological Sciences*. **2001**. *22(1)*. 46-47. (e) Fahn, S.; *Movement Disorder Society's 10th International Congress of Parkinson's Disease and Movement Disorders*. **2006**. Kyoto, Japan.

[51] (a) **IUPAC**: 2-[4-(2-methylpropyl)phenyl]propanoic acid. (b) Adams, S.; Nicholson, J.; Burrows, C.; Patented by The Boots Group. **1961**.

vivo. Further *in vivo* testing, however, revealed the existence of an isomerase (2-arylpropionyl-CoA epimerase) which converts (*R*)-ibuprofen to the active (*S*)-enantiomer[52]. Thus, the active principle is administered as a racemate avoiding the unnecessary chiral purification.

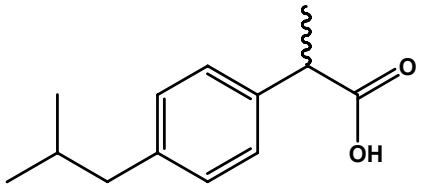


Figure 2.17: Ibuprofen.

2.2.1.3 Identical qualitative - different quantitative activity for each enantiomer

This is the case of both propranolol [Fig. 2.18] and dexfenfluramine [Fig. 2.19]. The first molecule, propranolol[53], is a well known non-selective beta blocker mainly used in the treatment of hypertension[54]. It is marketed as a racemate since both *S*- and *R*-enantiomers are active. The point is that the *S*-enantiomer -the eutomer- is about 40 times more effective than the *R*-enantiomer.

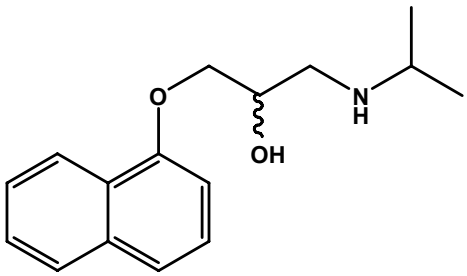


Figure 2.18: Propranolol.

[52] (a) Chen, C.S.; Shieh, W.R.; Lu, P.H.; Harriman S.; Chen, C.Y.; *Biochim Biophys Acta*. **1991**. *1078(3)*. 411-417. (b) Tracy, T.S.; Hall, S.D.; *Drug Metab Dispos*. **1992**. *20(2)*. 322-327. (c) Reichel, C.; Brugger, R.; Bang H.; Geisslinger, G.; Brune K.; *Mol Pharmacol*. **1997**. *51(4)*. 576-582.
[53] **IUPAC**: 1-(isopropylamino)-3-(naphthalen-1-yloxy)propan-2-ol.
[54] Rossi, S.; *Adelaide: Australian Medicines Handbook*. Australian Medicines Handbook Pty Ltd. **2006**. ISBN 0-9757919-2-3.

The second molecule in question, dexfenfluramine[55], is a serotoninergic anorectic drug marketed for the purposes of weight loss. In this compound, the eutomer, (+)-dexfenfluramine, is 4 times more active than the distomer, (-)-dexfenfluramine.

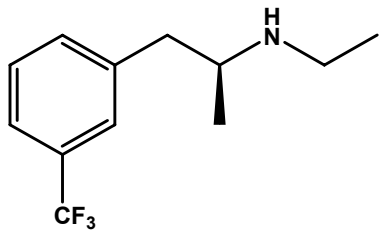


Figure 2.19: Dexfenfluramine.

2.2.1.4 Different qualitative activities for each enantiomer

There are many examples of molecules whose enantiomers show different activity both in a qualitative and quantitative sense.

Fluoxetine[56] (Prozac®) [Fig. 2.20] was initially marketed as a racemate until the phase II clinical trial for each enantiomer proved that *R*-fluoxetine was the eutomer responsible for the antidepressant effect[57], whilst *S*-fluoxetine was claimed to be the eutomer for the clinical treatment of migraine, although this point is under debate[58].

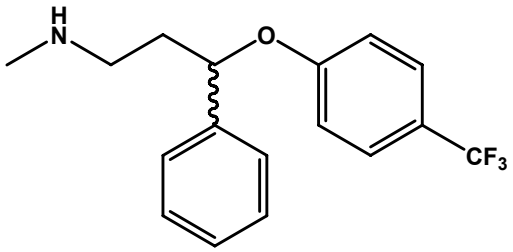


Figure 2.20: Fluoxetine.

[55] **IUPAC**: *N*-Ethyl-1-[3-(trifluoromethyl)phenyl]-propan-2-amine.
[56] (a) **IUPAC**: *N*-methyl-3-phenyl-3-[4-(trifluoromethyl)phenoxy]-propan-1-amine. (b) Wong, D.T.; Perry, K.W.; Bymaster, F.P.; *Nat. Rev. Drug Discov*. **2005**. *4(9)*. 764-774.
[57] (a) Carlsson, A.; Wong, D.T.; *Life Sci*. **1997**. *61(12)*. 1203-1203. (b) Wong, D.; Horng, J.; Bymaster, F.; Hauser, K.; Molloy, B.; *Life Sci*. **1974**. *15(3)*. 471-479. (c) Wong, D.T.; Bymaster, F.P.; Engleman, E.A.; *Life Sci*. **1995**. *57(5)*. 411-441. (d) Benfield, P.; Heel, R.C.; Lewis, S.P.; *Drugs*. **1986**. *32(6)*. 481-508.
[58] Steiner, T.J.; Ahmed, F.; Findley, L.J.; et al.; *Cephalalgia*. **1998**. *18(5)*. 283-286.

Propoxyphene also has two diastereomers: dextropropoxyphene[59] (Darvon®) [Fig. 2.21] is a mild opioid pain-killer and the eutomer acting as an analgesic[60]. Its diastereomer, levopropoxyphene[61] (Novrad®), appears to exert only an antitussive effect[62].

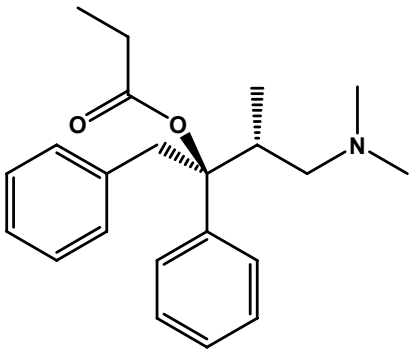


Figure 2.21: D-propoxyphene.

There are other sadly well known examples, like the non-steroidal anti-inflammatory drug benoxaprofen[63], or terodiline[64], the most commonly prescribed drug in Europe for the treatment of urinary urge incontinence[65], both of them presenting a pharmacologically active enantiomer and also a toxic one in the racemic mixture. Nevertheless, the most significant example is, undoubtedly, thalidomide.

2.2.1.5 The Thalidomide Disaster

Thalidomide[66] was a drug originally developed by the German pharmaceutical Group Grünenthal in 1957 [Fig. 2.22].

[59] **IUPAC**: [(2*R*,3*R*)-4-dimethylamino-3-methyl-1,2-diphenyl-butan-2-yl]propanoate.
[60] (a) Collins, S.L.; Edwards, J.E.; Moore, R.A.; McQuay, H.J.; *Cochrane Database Syst Rev*. **2000**. *(2)*. CD001440. (b) Moore, A.; Collins, S.; Carroll, D.; McQuay, H.; Edwards, J.; *Cochrane Database Syst Rev*. **2000**. *(2)*. CD001547. (c) Barkin, R.L.; Barkin, S.J.; Barkin, D.S.; *Am. J. Ther*. **2006**. *13(6)*. 534-542.
[61] **IUPAC**: [(2*R*,3*S*)-4-dimethylamino-3-methyl-1,2-diphenyl-butan-2-yl]propanoate.
[62] Wainer, I.W.; *Am. J. Hosp. Pharm*. **1992**. *49(9 Suppl 1)*. 4-8.
[63] (a) **IUPAC**: 2-[2-(4-chlorophenyl)-1,3-benzoxazol-5-yl]propanoic acid. (b) Marschall, E.; *Science*. **1985**. *229(4718)*. 1071-1071.
[64] (a) **IUPAC**: N-tert-butyl-4,4-di(phenyl)butan-2-amine (b) Connolly, M.J.; Astridge, P.S.; White, E.G.; Morley, C.A.; Cowan, J.C.; *Lancet*. **1991**. *338(8763)*. 344-345.
[65] Langtry, H.D.; McTavish, D.; *Drugs*. **1990**. *40(5)*. 748-761.
[66] (a) **IUPAC**: (±)-*N*-(2,6-Dioxo-3-piperidinyl)-1*H*-isoindol-1,3(2*H*)-dione (b) Mückler, H.; Patented by The Grünenthal Group. **1957**.

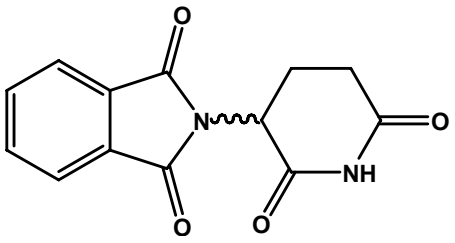


Figure 2.22: Thalidomide.

Marketed in Germany under the name Contergan® and all over the world in different other names, it was commonly prescribed as a sedative[67], hypnotic, and also as an antiemetic, to combat morning sickness and as an aid to help pregnant women sleep according to its *"unusual safety"*[68].

Nevertheless, studies undertaken by Lenz and McBride[69] around 1962 showed that thalidomide was the first highly teratogenic medicine discovered[70] and responsible for at least 8,000 birth defects in 46 countries[71]. From 1956 to 1962, approximately 10,000 children were born with severe malformations, including phocomelia, because their mothers had taken thalidomide during pregnancy[72].

Exhaustive research was carried out in the forthcoming years concluding that, because of the drug was a racemate, the *R*- enantiomer was effective against morning sickness, whilst the *S*- was teratogenic and responsible for the birth defects [Fig. 2.23]. Moreover, tests also proved that both enantiomers could interconvert *in vivo*[73]. Therefore administering only one enantiomer would not prevent the teratogenic effects in humans.

[67] (a) Wettstein, A.R.; Meagher, A.P.; Meagher, A.P.; *Lancet*. **1997**. *350(9089)*. 1445-1446. (b) Tseng, S.; Pak, G.; Washenik, K.; Pomeranz, M.K.; Shupack, J.L.; *J. Am. Acad. Derm*. **1996**. *35(6)*. 969-979.
[68] Koren, G.; Pastuszak, A.; Ito, S.; *N. Engl. J. Med*. **1998**. *338(16)*. 1128-1137.
[69] (a) Lenz, W.; Pfeiffer, R.A.; Kosenow, W.; Hayman, D.J.; *Lancet*. **1962**. *279(7219)*. 45-46. (b) Burley, D.M.; Lenz, W.; *Lancet*. **1962**. *279(7223)*. 271-272. (c) Knapp, K.; Lenz, W.; Nowack, E.; *Lancet*. **1962**. *280(7258)*. 725-725. (d) Mcbride, W.G.; Lenz, W.; Bignami, G.; Bovet, D.; Bovet-Nitti, F.; Rosnati, V.; Hobolth, N.; *Lancet*. **1962**. *280(7269)*. 1332-1334. (e) Lenz, W.; Maier, W.; *Lancet*. **1964**. *284(7369)*. 1124-1125. (f) Lenz, W.; *Medical Genetics*. University of Chicago Press. **1963**. (g) Lenz, W.; *Am. J. Dis. Child*. **1966**. *112*. 99-l06. (h) Lenz, W.; *Lecture given at the 1992 UNITH Congress*. **1992**.
[70] (a) Dally, A.; *Lancet*. **1998**. *351(9110)*. 1197-1199. (b) Vanchieri, C.; *Ann. Intern. Med*. **1997**. *127(10)*. 951-952.
[71] Florence, A.L.; *Brit. Med. J*. **1960**. *2(5217)*. 1954-1954.
[72] Bren, L.; *FDA Consumer, US Food and Drug Administration*. **2001**. *35(2)*. Cover Story.
[73] Eriksson, T.; Bjorkman, S.; Roth, B.; Fyge, A.; Hoglund, P.; *Chirality*. **1995**. *7(1)*. 44-52.

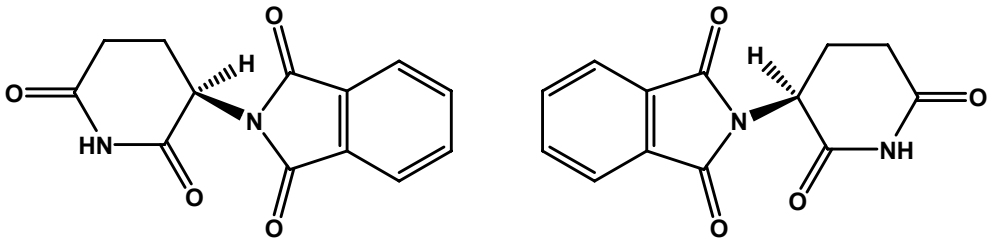


Figure 2.23: Enantiomers of Thalidomide. *R*-Thalidomide (left) sleep-inducing. *S*-Thalidomide (right) teratogenic.

2.2.1.6 Thalidomide: Consequences to Life and Economy

Grünenthal's medicine caused the greatest tragedy in the history of the German pharmaceutical industry and maybe in the world. Consequences to life and economy still persist today[74]: The criminal case opened on 27 May 1968 and closed on 18 December 1970 lasting about 2 years. After one of the longest and most complex trials up to that point in time in Germany, the company paid a voluntary sum of 114 million Deutschmarks into the *"Disabled Children's Relief Foundation"* and the federal government added another 100 million. Up to the end of 2007, a total sum of approximately 440 million euros has been paid and there are still approximately 2,800 people affected today in Germany by thalidomide who are entitled to benefits[75].

For this reason, chirality is nowadays one of the most important topics regarding organic chemistry, health and economics.

Those dramatic events deeply impressed researchers, inducing them to focus their interests into stereoselective syntheses and chiral separation techniques. Unfortunately, while chirality plays a main role in current organic chemistry, stereospatial properties induced by conformational changes hardly ever receive the same level of attention from chemists. Actually, this kind of studies falls commonly in the area of knowledge of biochemists, catching easily their attention due to their natural interest in macromolecules.

[74] Fusenig, A.; Compe, C.; Ramm H-J.; *1946-2006: 60 Years of Grünenthal*. Corporate Communication. Grünenthal GmbH, Aachen. **2006**.
[75] Bischoff, H.; Fusenig, A.; *The Thalidomide Tragedy -offprint-*. Grünenthal GmbH, Aachen. **2007**.

However, there are a few medical examples of highly relevant interest that prove the importance of molecular conformations. Probably, the better known one is that of the Creutzfeldt-Jakob disease because of its two epidemic outbreaks in Great Britain in the 90's.

2.2.2 Structure-Activity Relationships II: Conformation. Creutzfeldt-Jakob Disease

There are several kinds of incurable degenerative neurological disorders classified in the type of *"transmissible spongiform encephalopathy"*. Brain diseases include Creutzfeldt-Jakob disease, Gerstmann-Sträussler-Scheinker syndrome, fatal familial insomnia, sporadic fatal insomnia, and kuru in humans, as well as bovine spongiform encephalopathy (commonly known as mad cow disease), chronic wasting disease in elk and deer, and scrapie in sheep, transmissible mink encephalopathy, feline spongiform encephalopathy, and exotic ungulate encephalopathy, in animals.

The better known one in this group is the Creutzfeldt-Jakob disease[76] (CJD). It is a very rare, incurable, and ultimately fatal disease caused by prions.

2.2.2.1 Prions

Thirty-five years ago, little was known about the causes of neurodegenerative disorders. The current explanation, a controversial theory still under debate[77], was proposed by Dr. Stanley B. Prusiner who claimed for the existence of an agent called *"Prion"*[78] - proteinacious infectious agent- that is ultimately responsible for the transmission of those diseases. He was awarded the Nobel Prize in Physiology or Medicine in 1997 for his prion research [Fig. 2.24].

[76] (a) Balter, M.; *Science*. **2000**. *289(5484)*. 1452-1454. (b) Ghani, A.C.; Ferguson, N.M.; Donnelly, C.A.; Anderson, R.M.; *Nature*. **2000**. *406(6796)*. 583-584.
[77] (a) Prusiner, S.B.; *Development of the prion concept. In: Prusiner SB, ed. Prion biology and diseases*. Cold Spring Harbour, N.Y. Cold Spring Harbour Laboratory Press. **1999**. 67-112. (b) Manuelidis* L.; Yu, Z-X.; Barquero, N.; Mullins, B.; *Proc. Nat. Acad. Sci. USA*. **2007**. *104(6)*. 1965-1970.
[78] (a) Prusiner, S.B.; *N. Engl. J. Med*. **2001**. *344(20)*. 1516-1526. (b) Prusiner S.B.; *Proc. Nat. Acad. Sci. USA*. **1998**. *95(23)*. 13363-13383. (c) Prusiner, S.B.; *N. Engl. J. Med*. **1984**. *310(10)*. 661-663. (d) Prusiner, S.B.; *Brain Pathol*. **1998**. *8(3)*. 499-513. (e) Prusiner, S.B.; *Science*. **1982**. *216(4542)*. 136-144. (f) Prusiner, S.B.; *Science*. **1991**. *252(5012)*. 1515-1522.

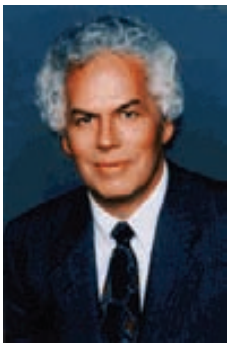


Figure 2.24: Dr. Stanley B. Prusiner

2.2.2.2 Prions and the Creutzfeldt-Jakob Disease

The protein that CJD prions are made of is found throughout the body, even in healthy people and animals. It has 209 amino acids (in humans), one disulfide bond, a molecular weight of 35-36 kDa and a mainly α-helical structure. Anyway, several other topological forms exist[79].

Prions are hypothesized to infect and propagate by changing their conformation[80], refolding abnormally into a structure which is able to convert normal molecules of the protein PrP^{C} into the abnormally structured form PrP^{Sc} [Fig. 2.25]. That is, the infected protein is able to act as a catalyst for the transformation.

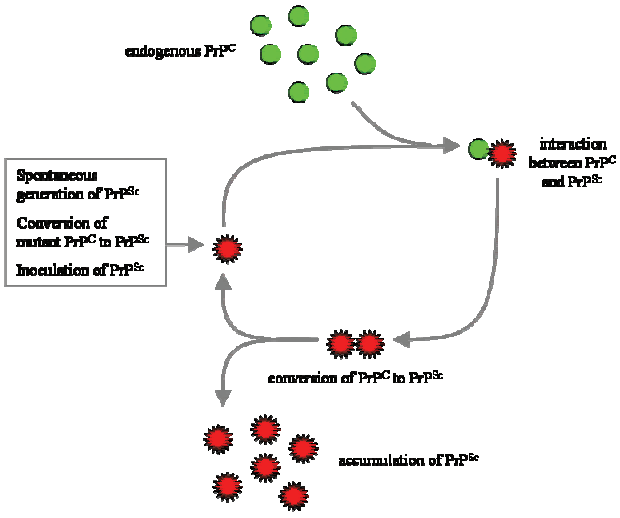


Figure 2.25: Catalytic Process. The infected protein catalysing the transformation.

[79] Hegde, R.S.; Mastrianni, J.A.; Scott, M.R.; Defea, K.A.; Tremblay, P.; Torchia, M.; DeArmond, S.J.; Prusiner, S.B.; Lingappa, V.R.; *Science*. **1998**. *279(5352)*. 827-834.
[80] (a) Harris, D.A.; True, H.K.; *Neuron*. **2006**. *50(3)*. 353-357. (b) Ronga, L.; Tizzano, B.; Palladino, P.; Ragone, R.; Urso, E.; Maffia, M.; Ruvo, M.; Benedetti, E.; Rossi, F.; *Chem. Biol. Drug Des*. **2006**. *68(3)*. 139-147.

The exact 3D structure of PrP^{Sc} is yet not known but there is increased β-sheet content in the diseased form of the molecule replacing normal areas of α-helix[81] [Fig. 2.26].

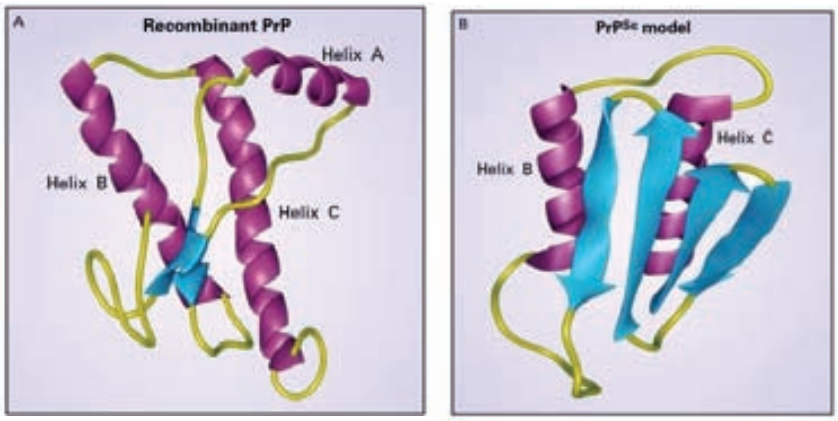


Figure 2.26: PrP Conformations: (a) Ordinary and (b) Toxic.

All known prions, as a consequence of the reduced solubility of the PrP^{Sc} isoform, induce the formation of an amyloid fold in which the protein clots into an aggregate consisting of tightly packed β-sheets. This altered structure is extremely stable, resistant to proteases, and accumulates in infected tissue, causing cell death, tissue damage, and death[82].

2.2.2.3 Neurodegenerative Diseases in the Future. Consequences to Life and Economy

On the one hand, the importance of neurodegenerative diseases and brain disorders concerning abnormal processing or folding of neuronal proteins was clearly stated by Prusiner in his Shattuck Lecture in 2001, focusing on two of them; Parkinson and Alzheimer:

"With the increase in life expectancy, there has been concern about the incidence of Alzheimer's and Parkinson's diseases. Among persons who are 60 years old, the prevalence of Alzheimer's disease is approximately 1 in 10,000, but among those who

[81] (a) Pan, K.M.; Baldwin, M.; Nguyen, J.; Gasset, M.; Serban, A.; Groth, D.; Mehlhorn, I.; Huang, Z.; Fletterick, R.J.; Cohen, F.E.; et al. *Proc. Nat. Acad. Sci. USA*. **1993**. *90(23)*. 10962-10966. (b) Riek, R.; Hornemann, S.; Wider, G.; Billeter, M.; Glockshuber, R.; Wüthrich, K.; *Nature*. **1996**. *382(6587)*. 180-182.
[82] Dobson, C.M.; *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. **2001**. *356(1406)*. 133-145.

are 85 years old, it is greater than 1 in 3. These data suggest that by 2025, there will be more than 10 million cases of Alzheimer's disease in the United States, and by 2050, the number will approach 20 million. The annual cost associated with Alzheimer's disease in the United States is estimated at $200 billion. Age is also the most important risk factor for Parkinson's disease. Nearly 50 percent of persons who are 85 years old also have at least one symptom or sign of parkinsonism." [83]

On the other hand, there is still a group of brain disorders –some of them with low statistical incidence- also considered to be related to conformational diseases: amyotrophic lateral sclerosis, frontotemporal dementia, Huntington's disease, and spinocerebellar ataxias [Fig. 2.27].

Disease	Nº of Cases	Nº per 100,000 (*)
Prion disease	400	<1
Alzheimer's disease	4,000,000	1,450
Parkinson's disease	1,000,000	360
Frontotemporal dementia	40,000	14
Pick's disease	5,000	2
Progressive supranuclear palsy	15,000	5
Amyotrophic lateral sclerosis	20,000	7
Huntington's disease	30,000	11
Spinocerebellar ataxias	12,000	4

** Data are based on a population of aproximately 275 million in 2000*

Figure 2.27: Prevalence of Neurodegenerative Diseases in the United States in 2000. Prusiner. *Shattuck Lecture.*

Therefore, we think that these data clearly suggest that, while chirality is important, it is also worth paying more attention to conformational studies due to their indisputable relevance.

2.3 Molecular Modelling and Conformational Analysis

As previously shown, the *Folding@home* project and recent discoveries in neurodegenerative diseases claim for the importance of conformational studies that

[83] Presented as the 110th Shattuck Lecture to the Annual Meeting of the Massachusetts Medical Society, Boston, May 20th, **2000**. From the Institute for Neurodegenerative Diseases and the Departments of Neurology and of Biochemistry and Biophysics, University of California, San Francisco. [Prusiner, S.B.; *N. Engl. J. Med*. **2001**. *344(20)*. 1516-1526.]

allow a better understanding of the folding mechanism. Studies that, working in parallel with other disciplines, could help to find medical and pharmaceutical solutions to problems of high interest for our Society.

Conformational Analysis, the main theme within the present research work, is part of *Molecular Modelling*[84]; a multidisciplinary science which involves Chemistry, Biochemistry, Mathematics, and Computer Sciences. In the words of Andrew R. Leach[84]: *"Today, molecular modelling is invariably associated with computer modelling"* which make sense according to *Levinthal's Paradox Corollary*. Therefore, it is not difficult to realise that computational chemistry development has been closely linked to computer science and algorithmic revolutions.

2.3.1 Algorithms, Storing and Analysing; "Data Mining"

During the last century there have been mainly three Eras in computer science development according to the way the problem was approached: How can I solve my problem? How can I store my results? And, how can I analyse my results?

2.3.1.1 How can I solve my problem?

It was almost at the beginning and focused on the logical and mathematical aspects of computing. It might be considered the ***Algorithmic Era*** and names like Alan Mathison Turing (1912–1954), John von Neumann (1903–1957) and Nicholas Constantine Metropolis (1915–1999) were widely recognised.

The war and the forthcoming years witnessed an outburst of new ideas most of them related to logic, cryptography, and atomic physics. But this was also the time when people began to wonder about the nature of problems that can or cannot be solved by mathematical logics, and therefore, computational means; and about artificial intelligence, or whether it will be possible to say that a machine is conscious and can

[84] (a) Leach, A.R.; *Molecular Modelling: Principles and Applications*. 2nd Ed. Prentice Hall. Pearson Education Limited. **2001**. ISBN 0-582-38210-6. (b) Goodman, J.M.; *Chemical Applications of Molecular Modelling*. Cambridge, Royal Society of Chemistry. **1998**. ISBN 0-85404-579-1.

think. It seems quite a bit *naïve* or philosophical, but this was a crucial point in computer science. Until this time, problems had been more or less divided into these two categories: continuous and discrete.

2.3.1.1.1 The continuous problem

The continuous problems are mainly those of classical physics and mathematics in which a continuous function/functional is studied, trying to find an extreme that optimizes it. The function/functional can be single or multivariable, and can also be constrained by boundary conditions and/or restrictions. I.e. finding the minimum side area of a cylindrical recipient where volume is already fixed, or calculating the Earth's planetary orbit [Fig. 2.28].

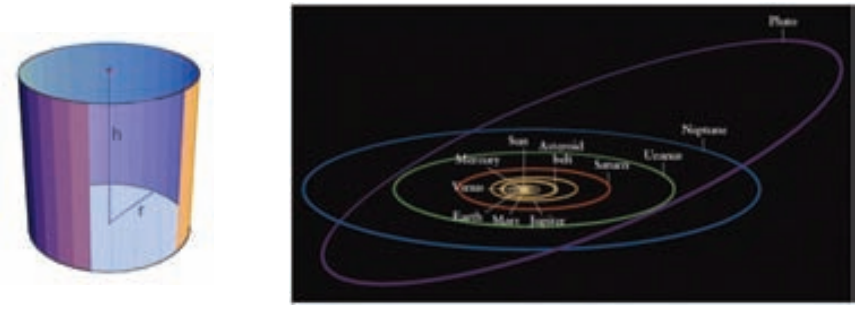


Figure 2.28: Classical Problems: (a) For a given volume "V" (restriction), find the best pair of "h" and "r" values which minimises cylindrical side area. (b) Planetary Orbits in the Solar System.

Differential Calculus has been solving problems of this nature for nearly 300 years: Sir Isaac Newton (1643-1727), Leonhard Paul Euler (1707-1783), Joseph-Louis de Lagrange (1736-1813) and Sir William Rowan Hamilton (1805-1865) developed this part of the mathematical-physics until its furthest edges[85], while computer sciences along the 20th century mainly "translated" analytical solutions into a computer language. Several already considered *classical algorithms*[86] were proposed during these years and

[85] (a) Goldstein H.; Poole C.P.; Safko J.L.; *Classical Mechanics*. 3rd Ed. Addison Wesley. **2001**. ISBN-13 978-0201657029. (b) Hand, L.N.; Finch, J.D.; *Analytical Mechanics*. Cambridge University Press. **1998**. ISBN 0-521-57572-9.

[86] (a) Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P.; *Numerical Recipes: The Art of Scientific Computing*. 3rd Ed. Cambridge University Press. **2007**. ISBN-10: 0521880688 (b) Knuth, D.E.; *The Art of Computer Programming*. Vol. 1: *Fundamental Algorithms*. 3rd Ed. **1997**. ISBN 0-201-89683-4. Vol. 2: *Seminumerical Algorithms*. 3rd Ed. **1997**. ISBN 0-201-89684-2. Vol. 3: *Sorting and Searching*.

have been extensively used in computational chemistry mainly in the area of ***Geometrical-Energetic Optimization***.

2.3.1.1.2 The discrete problems

The discrete problems are those in which either it is not possible to model the problem as a continuous function/functional or a classical formalisation is impracticable. They are commonly referred as *combinatorial problems*, and the set of feasible solutions is discrete or can be reduced to a discrete one. In these cases, the goal is to find the best possible solution.

Typical examples are the *travelling salesman problem*[87], consisting of, given a number of cities and the costs of travelling from any city to any other city, to guess the least-cost round-trip route that visits each city exactly once and then returns to the starting city. *The eight queens puzzle*[88], where the problem is solving how to arrange eight chess queens on a chessboard in a way that none of them is able to capture any other. *The knapsack problem*[89], which derives its name from the maximization problem of the best choice of essentials that can fit into one bag. Or the *minimum spanning tree*[90], which is, given a connected, undirected graph, to find a subgraph which is a tree and connects all the vertices together weighting less than, or equal to, the weight of every other spanning tree [Fig. 2.29].

2nd Ed. **1998**. ISBN 0-201-89685-0. Vol. 4: *Combinatorial Algorithms* (in preparation). Addison-Wesley Professional.

[87] (a) Applegate, D.L.; Bixby, R.E.; Chvátal, V.; Cook, W.J.; *The Traveling Salesman Problem: A Computational Study*. Princeton University Press. **2006**. ISBN 978-0-691-12993-8. (b) Lawler, E.L.; Lenstra, J.K.; Rinnooy Kan, A.H.G.; Shmoys, D.B.; *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. John Wiley & Sons. **1985**. ISBN 0-471-90413-9.

[88] Dahl, O.-J.; Dijkstra, E.W.; Hoare, C.A.R.; *Structured Programming*. Academic Press, London. **1972**. ISBN 0-12-200550-3.

[89] (a) Martello, S.; Toth P.; *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons. **1990**. ISBN 0-471-92420-2. (b) Kellerer, H.; Pferschy, U.; Pisinger, D.; *Knapsack Problems*. Springer Verlag. **2005**. ISBN 3-540-40286-1.

[90] Pettie, S.; Ramachandran, V.; *JACM*. **2002**. *49(1)*.16-34.

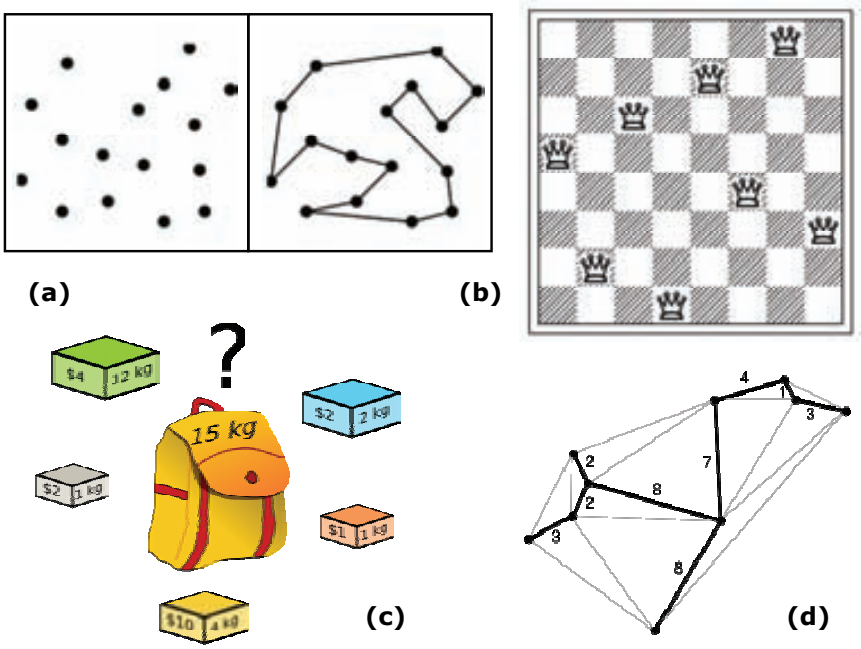


Figure 2.29: Collected examples of Combinatorial Problems: (a) Travelling Salesman Problem, (b) The Eight Queens Puzzle, (c) The Knapsack Problem, and (d) Minimum Spanning Tree Problem.

Continuous algorithms were unable to solve these new problems where the conformational space had to be extensively explored because they were not designed for this purpose. First attempts using *metropolis-montecarlo*[91] helped significantly and, in fact, were a real breakthrough in the classical programming think-tank. Nevertheless, the turning point in the algorithmic era was the crucial contribution made by Holland[92] when he published his ground-breaking book on *Genetic Algorithms* based upon the *Darwinian Evolutionary Theories*[93]. This book, offering a new solution for the so called *Combinatorial Optimization Problems*, encouraged new scientist in the field to investigate and move towards *metaheuristic approaches*[94].

It was more like a fever, and in less than a decade, programmers all over the world "discovered" that Nature itself was able to solve these problems. Mimicking Nature

[91] (a) Ulam, S.; Richtmyer, R.D.; von Neumann, J.; Los Alamos Scientific Laboratory Report. **1947**. *LAMS-551*. (b) Metropolis, N.; Ulam, S.; *J. Am. Statist. Ass.* **1949**. *44(247)*. 335-341. (c) Chang, G.; Guida, W.C.; Still, W.C.; *J. Am. Chem. Soc*. **1989**. *111(12)*. 4379-4386. (d) Chang, G.; Guida, W.C.; Still, W.C.; *J. Am. Chem. Soc*. **1990**. *112(4)*. 1419-1427.

[92] (a) Holland, J.H.; *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA. **1975**. Reprint edition **1992**. ISBN-10 0-262-58111-6. (b) Goldberg, D.E.; *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. **1989**. ISBN-13 978-0201157673.

[93] (a) Darwin, C.R.; *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray. 1st edition. 1st issue. **1859**. (b) http://darwin-online.org.uk

[94] Blum, C.; Roli, A.; *ACM Computing Surveys*. **2003**. *35(3)*. 268-308.

became the new paradigm[95] in computer science, and suddenly, whoever would still dare to speak of the prosperity of classical methods was automatically regarded as old-fashioned and outdated, and in the end, almost "extinct".

A whole family of non continuous algorithms are nowadays in use: *Evolutionary Algorithms*[93] were developed after Holland (i.e. *Genetic Algorithms*[92], *Genetic Programming*[96], *Evolutionary Programming*[97], *Evolution Strategies*[98] and *Learning Classifier Systems*), but also *animal-behaviour based algorithms* (i.e. *ant colony optimization*[99], *swarm intelligence*[100]…), non *nature-inspired algorithms* (i.e. *tabu search*[101]), and also several variants for the *Montecarlo*[91] (i.e. *Simulated Annealing*[102], *Stochastic Optimization*[103]…).

The reason for this computational survey is that ***Conformational Search is also a problem of combinatorial optimization***. Most of the algorithms are still being used, so there is not a "best" conformational search algorithm and the choice usually depends on the size of your molecule, and the topology. Briefly, our experience says that, for small molecules, *Metropolis-Montecarlo* is usually a good option, nevertheless, as the size of

[95] Kuhn, T.; *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago. **1962** y **1969**. ISBN 0226458083.
[96] Banzhaf, W.; *Genetic Programming and Evolvable Machines*. Springer, US. 10701. **2002**. ISSN. 1389-2576.
[97] (a) Fogel, L.J.; Walsh, M.J.; *Artificial Intelligence through Simulated Evolution*. John Wiley & Sons, New York. **1966**. ISBN-13 978-0471265160. (b) Fogel, D.B.; *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence*. IEEE Press. Piscataway. NJ. Wiley-IEEE Press; 2nd Ed. **1999**. ISBN-13 978-0780353794.
[98] (a) Rechenberg, I.; *Doctoral Thesis*. *Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution (in German)*. **1971**. Reprinted by Fromman-Holzboog in **1973**. (b) Beyer, H.-G.; Schwefel. H.-P.; *Journal Natural Computing*. **2002**. *1(1)*. 3-52. (c) Beyer, H.-G.; *The Theory of Evolution Strategies*. Springer. **2001**. ISBN-13 978-3540672975.
[99] (a) Dorigo, M.; *Doctoral Thesis (in Italian)*. DEI. Politecnico di Milano, Italy. **1992**. (b) Dorigo, M.; Maniezzo V.; Colorni, A.; *IEEE Transactions on Systems, Man, and Cybernetics–Part B*. **1996**. *26(1)*. 29-41. (c) Dorigo M.; Gambardella, L.M.; *IEEE Transactions on Evolutionary Computation*. **1997**. *1(1)*. 53-66. (d) Dorigo, M.; Di Caro, G.; Gambardella, L.M.; *Artificial Life*. **1999**. *5(2)*. 137–172. (e) Dorigo, M.; Stützle, T.; *Ant Colony Optimization*. MIT Press. **2004**. ISBN 0-262-04219-3.
[100] (a) Kennedy, J.; Eberhart, R.; *"Particle swarm optimization". Proc. of the IEEE Int. Conf. on Neural Networks. Piscataway, NJ*. **1995**. 1942–1948. (b) Bonabeau, E.; Dorigo, M.; Theraulaz, G.; *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press. **1999**. ISBN 0-19-513159-2. (c) Kennedy, J.; Eberhart, R.C.; Yuhui, S.; Shi, Y.; *Swarm Intelligence*. Morgan Kaufmann Publisher, San Francisco, CA. **2001**. ISBN 1-55860-595-9.
[101] (a) Glover, F.; *Decision Sciences*. **1977**. *8(1)*. 156-166. (b) Glover, F.; *Comput. Oper. Res*. **1986**. *13(5)*. 533-549. (c) Glover, F.; Laguna, M.; *Tabu Search*. Kluwer Academia Publishers. **1997**. ISBN 0-7923-8187-4.
[102] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.; *Science*. **1983**. *220(4598)*. 671-680.
[103] (a) Robbins, H.; Monro, S.; *Annals of Mathematical Statistics*. **1951**. *22(3)*. 400-407. (b) J. Kiefer; J. Wolfowitz. *Annals of Mathematical Statistics*. **1952**. *23(3)*. 462-466. (c) Spall, J.C.; *Introduction to Stochastic Search and Optimization*. John Wiley & Sons. **2003**. ISBN-13 978-0471330523.

your system grows, *Genetic-Algorithms* and *Simulated Annealing* become more and more effective.

2.3.1.2 How can I store my Results?

It seemed that the "problem" regarding algorithms was more or less put aside by the end of the 80's in the assumption that it was almost solved. Improvement in microprocessors following *Moore's Law* -doubling computational power almost every 18 months[104]- had led into a significant reduction in processing time [Fig. 2.30].

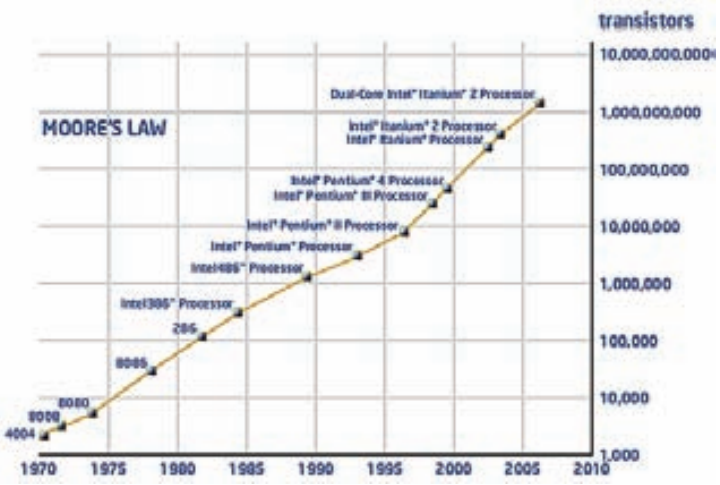


Figure 2.30. ***Moore's Law*****. INTEL©.**

Then, computational researchers began to generate extremely large amounts of data in the hope of getting better results. But then, a new problem arose: how can I store all my results? This was the beginning of the ***Storing-Devices Era***. Technology rapidly evolved and again advances in magnetoresistance and magnetic surfaces[105], seemed to work in favour sustaining *Kryder's Law*, which states that the current rate of increase in hard drive capacity is roughly similar to the rate of increase in transistor count[106] (recent trends show that this rate has been maintained into 2007) [Fig. 2.31].

[104] (a) Moore, G.E.; *Electronics*. **1965**. *38(8)*. 114-117. (b) Moore, G.E.; *Excerpts of A conversation with Gordon Moore: Moore's Law*. Video Transcript. INTEL©.
[105] Ertl, G.; Nobel Prize in Chemistry. **2007**. *"For his studies of chemical processes on solid surfaces"*.
[106] Kryder, M.; *Scientific American*. **2005**. *August*.

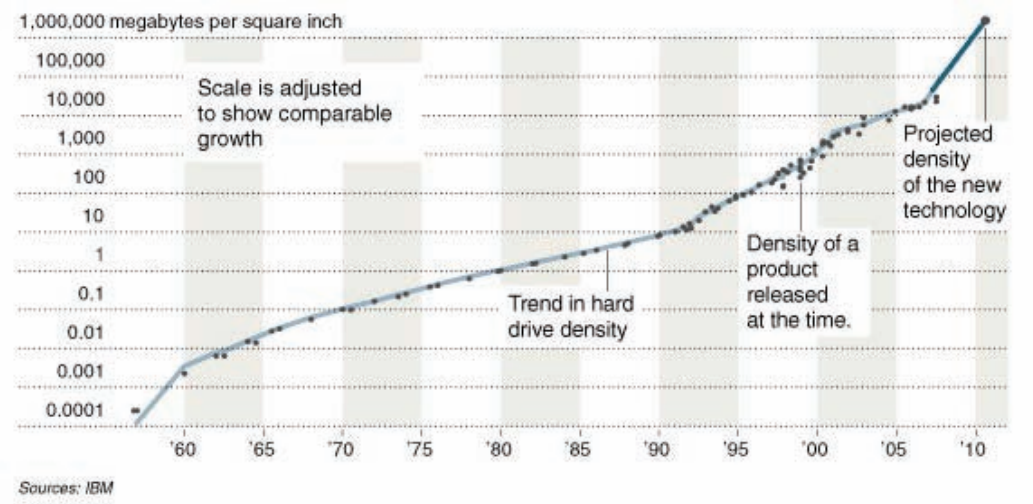


Figure 2.31: ***Kryder's Law*****. IBM©.**

Storing space is really important for conformational studies. Levinthal's Paradox exposed that the amount of possible conformations available for a given polypeptide chain is really large, and this information must be analysed to extract useful conclusions.

2.3.1.3 How can I analyze my Results?

In principle, it seemed that there would be no more drawbacks. But again a new problem emerged around the mid 90's. Faster and faster computers, usually produce data files the size of hundreds of Gigabytes of results. Output files generated after calculations lasting months, stored in bigger and even bigger storing-devices, have to be analysed. But it is not possible to extract information easily from such collections of figures. This means new problems, and also the beginning of the third Era: ***The Data-Mining Era***.

Currently, not only computational chemistry manages large data files. Government agencies, telecommunications companies, mass media, industries, scientific laboratories, hospitals, financial and insurance corporations, internet servers, searching engines, and many others administer databases that store millions, or maybe billions, of different data types. Searching, sorting, accessing, filtering, finding patterns, and in general, extracting useful information from collections whose size is immense has

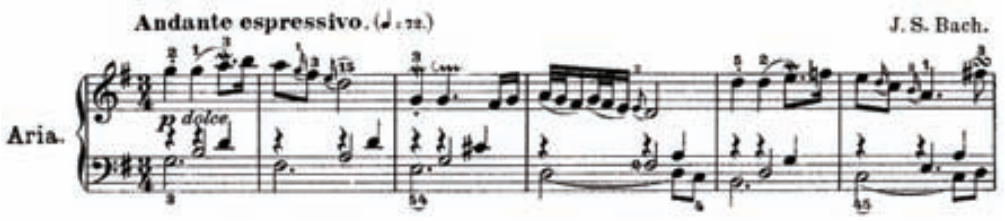
become in the last 15 years a mayor challenge and a new extremely useful branch in computational science: *Data Mining*.

Data Mining has been described in different ways: the nontrivial extraction of implicit, previously unknown, and potentially useful information from data[107], the science of extracting useful information from large data sets or databases[108], and finally in relation to Enterprise Resource Planning, Data Mining is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making[109].

Development of SQL, the *Structured Query Language*[110], (currently a standard in Data Bases), combined with implementation of learning classifier systems, searching algorithms (*neural networks*[111]) and advanced statistical techniques as multivariate analysis (*principal component analysis*[112], *factor analysis*[113]…) or classification by *clustering*[114] are today common tools in any area of knowledge like Computational Chemistry, and therefore, in Conformational Analysis.

[107] Frawley W.; Piatetsky-Shapiro G.; Mattheus C.; *AI Magazine*. **1992**. *13(3)*. 213-228.
[108] Hand, D.; Mannila, H.; Smyth, P.; *Principles of Data Mining*. MIT Press, Cambridge, MA. **2001**. ISBN 0-262-08290-X.
[109] Monk, E.; Wagner B.; *Concepts in Enterprise Resource Planning*, 2nd Ed. Thomson Course Technology, Boston, MA. **2006**. ISBN 0-619-21663-8.
[110] (a) Codd, E.F; *Communications of the ACM*. **1970**. *13(6)*. 377-387. (b) Codd, E.F.; *The Relational Model for Database Management, Version 2*. Addison Wesley Publishing Company. **1990**. ISBN 0-201-14192-2.
[111] Isasi Viñuela, P.; Galván León, I.M.; *Redes de Neuronas Artificiales: Un Enfoque Práctico*. Pearson Prentice Hall. **2004**. ISBN 84-205-4025-0.
[112] (a) Pearson, K.; *Phil. Mag.* **1901**. *2(6)*. 559-572. (b) Jolliffe I.T.; *Principal Component Analysis, Series: Springer Series in Statistics, 2nd Ed*. Springer. NY. XXIX. 487. **2002**. ISBN 978-0-387-95442-4
[113] Gorsuch, R.L.; *Factor Analysis*. 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum. **1983**.
[114] (a) Sokol, R.R.; *Clustering and Classification: Background and Current Directions*. in J. Van Ryzin, Classification and Clustering, Academic Press, New York, **1977**. ISBN 0127142509. (b) Zupan, J.; *Clustering of Data. Chapter 7. Algorithms for Chemists*. John Wiley and Sons, New York. **1989**. ISBN 0-471-92173-4.

CYCLODEXTRINS



Johann Sebastian Bach
Goldberg Variations BWV 988
Aria
(1741)

"Cyclic forms are everywhere..."

3 CYCLODEXTRINS

3.1 Introduction

Coming back to the point; once the overview of the whole work has been outlined in the Introduction -explaining the importance of conformational properties of macromolecules and the tools surrounding this sort of research- this is time to remember again our objectives.

Briefly, as said in Chapter I, our interests were set, on the one hand, to conformational search, and on the other hand, to conformational space exploration (Molecular Dynamics). Once the problem was identified we suggested a new methodological approach to give new answers and study macromolecular systems.

Therefore, on the basis of the precedents, we are now dealing with a ***methodological part*** -that will be explained in the next chapter and also in the appendix in chapter IX- and also with a ***macromolecular part*** that relates the set of molecules which will be used as a benchmark to test the methodology. This set will be explained within the current chapter.

Eventually, the choice of this family of sugars is not by hazard. Historically, our research group has been quite fond of cyclodextrins and their derivatives, and several thesis and papers have been entirely devoted to this field creating the necessary background in the area and establishing the theoretical basis and the *"know-how"* in computer science to afford this challenge.

3.2 A Survey in Cyclodextrins

Cyclodextrins (CD), also known as cycloamyloses (CA), make up a family of cyclic oligosaccharides composed of 5 or more α-D-glucopyranoside units linked 1→4. They contain a number of glucose monomers ranging from 5 to more than 100 units in a ring[115] although the most frequent ones are those considered "native" or "common", having 6, 7, and 8 units [Fig. 3.1].

[115] Takaha, T.; Yanase, M.; Takata, S.; Okada, S.; Smith, S.M.; *J. Biol. Chem*. **1996**. *271(6)*. 2902-2908.

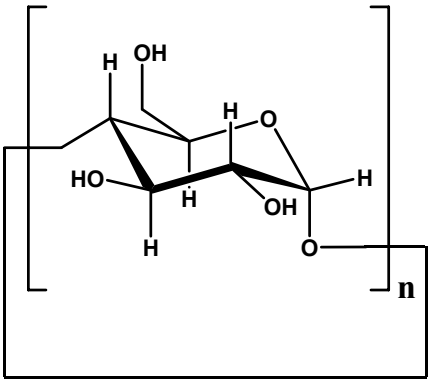


Figure 3.1: Schematic depiction of a Cyclodextrin.
The unit is a monomer of glucose that repeats "n" times.

Displaying amphoteric properties is the principal feature of this macromolecular family due to both the lipophilic central cavity and the hydrophilic outer surface, which allow them to form stable inclusion complexes [Fig. 3.2].

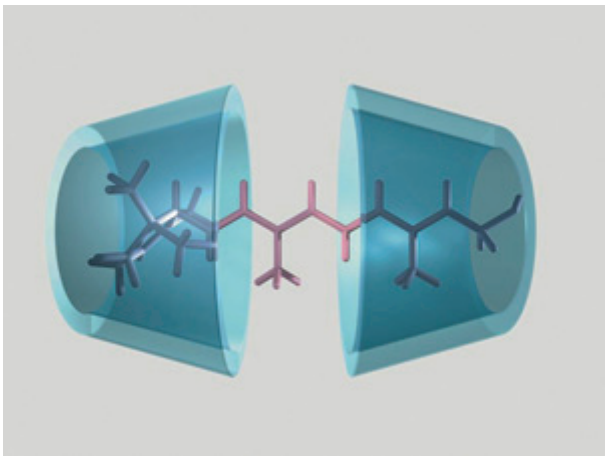


Figure 3.2: Example of host-guest complex:
Two cyclodextrins are encapsulating a substrate.

From the supramolecular viewpoint, these molecules can be topologically represented as *"toroids"*, bearing respectively secondary and primary hydroxyl groups exposed to the solvent, both on the larger and the smaller openings of the macroring [Fig. 3.3].

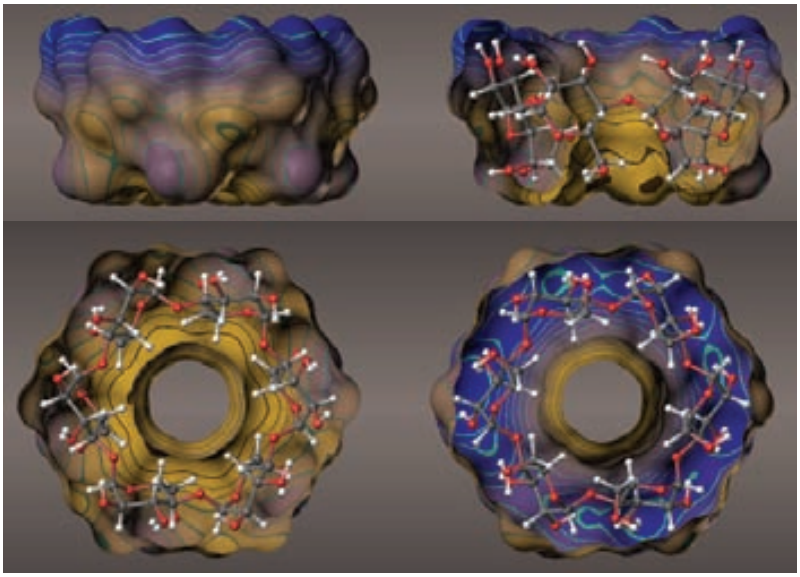


Figure 3.3: Upper, side and "internal" view of a α-cyclodextrin[116]. The toroidal shape and the inner cavity are easily recognisable. Hydroxyl groups oriented to the solvent are shown.

3.3 Brief History of Cyclodextrins

Cyclodextrins were firstly described in 1891 by A. Villiers[117], who named them *cellulosines*. A few years later, while studying food decay, F. Schardinger identified the three naturally occurring ones[118] α-, β-, and γ- (6, 7 and 8 glucoses) [Fig. 3.4], and soon after, during the period of time between 1911 and 1935, Pringsheim became the leading researcher in this area demonstrating that cyclodextrins formed stable aqueous complexes with many other chemical substances.

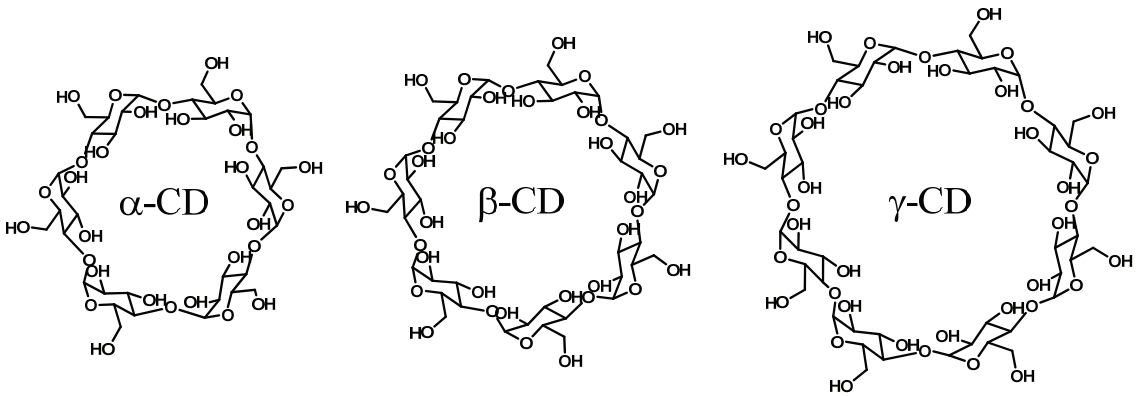


Figure 3.4: The three native cyclodextrins identified by Schardinger.

[116] Image source: Lichtenthaler, F.W.; Immel, S.; *Tetrahedron Asymmetry*. **1994**. *5(11)*. 2045-2060.
[117] Villiers, M.A.; *Compt. Rend. Fr. Acad. Sci*. **1891**. *112*. 435-438.
[118] Schardinger, F.; *Zentr. Bacteriol. Parasitenk. Abt. II*. **1911**. *29*. 188-197.

In 1935, Freudenberg and Jacobi developed a feasible purification method for the three natural cyclodextrins[119]. Soon after, in 1936, again Freudenberg and co-workers proposed a structural molecular model[120] consisting of a cyclic arrangement of units which was confirmed in further studies[121]. Few years later, around 1948, the first hint on the existence of rings containing a number of glucoses above 8 units was published[122] and also new discoveries in this area were made during the next decade, claiming for the existence of cyclodextrins made up of 9, 10, 11 and 12 glucose units[123].

By the mid 1970's, each of the native cyclodextrins and some large ones in the family had already been structurally and chemically characterized[124]. Extensive work on the subject was conducted from 1970 onward by Szejtli[125] and other authors exploring encapsulation by cyclodextrins and their derivatives in search for industrial and pharmaceutical applications. Finally, important advances made last decade in spectroscopic techniques allowed molecular characterization of giant cyclodextrins, proving their existence unambiguously[126].

3.4 Synthesis and Large Scale Production

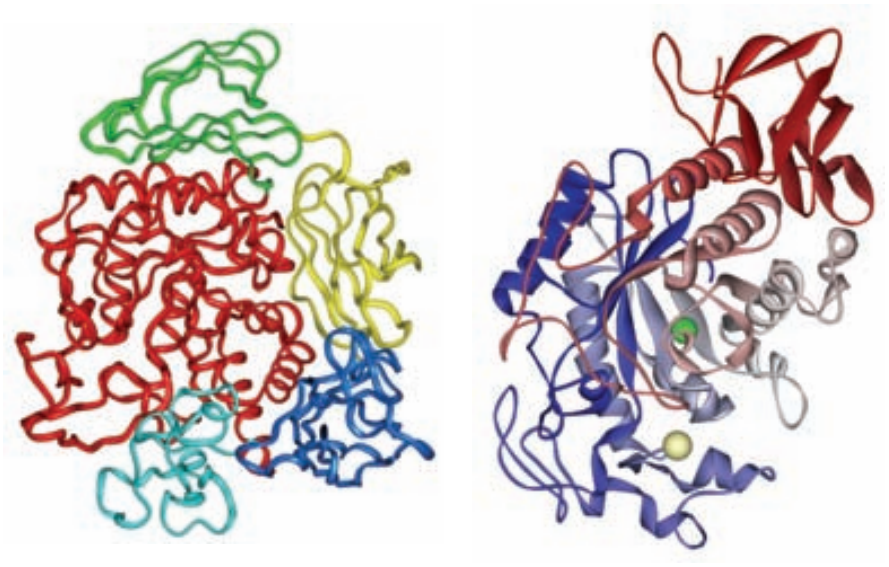
It took its time to introduce cyclodextrins in the chemical industry, probably because of their extremely high production expenses. Fortunately, although there are several laboratory scale preparative synthetic paths, cyclodextrins are nowadays produced at large scale from starch by means of a worthwhile enzymatic conversion[127].

---

[119] Freudenberg, K.; Jacobi, R.; *Ann. Chem*. **1935**. *518*. 102-108.
[120] Freudenberg, K.; Blomqvist, G.; Ewald, L.; Soff, K.; *Chem. Ber*. **1936**. *69B*. 1258-1266.
[121] French, D.; *Adv. Carbohydr. Chem*. **1957**. *12*. 189-260.
[122] Freudenberg, K.; Cramer, F.; *Z. Naturforsch*. **1948**. *3b*. 464.
[123] (a) Pulley, A.O.; French, D.; *Biochem. Biophys. Res. Commun*. **1961**. *5*. 11-15 (b) French, D.; Pulley, A.O.; Effenberger, J.A.; Rougvie, M.A.; Abdullah, M.; *Arch. Biochem. Biophys*. **1965**. *111(1)*. 153-160. (c) Endo, T.; Ueda, H.; Kobayashi, S.; Nagai, T.; *Carbohydrate Research*. **1995**. *269(2)*. 369-373. (d) Ueda, H.; Endo, T.; Nagase, H.; Kobayashi, S.; Nagai, T.; *J. Incl. Phenom. Macrocycl. Chem*. **1996**. *25(1-3)*. 17-20.
[124] Schneider, H.J.; Hacket, F.; Rüdiger, V.; *Chem. Rev*. **1998**. *98(5)*. 1755-1785.
[125] Szejtli, J.; *Cyclodextrin Technology*. Kluwer Academic Publishers: Dordrecht. **1988**.
[126] (a) Jacob, J.; Gessler, K.; Hoffmann, D.; Sanbe, H.; Koizumi, K.; Smith, S.M.; Takaha, T.; Saenger, W.; *Carbohydrate Research*. **1999**. *322(3-4)*. 228-246. (b) Nimz, O.; Gessler, K.; Usón, I.; Saenger, W.; *Carbohydrate Research*. **2001**. *336(2)*. 141-153. (c) Nimz, O.; Gessler, K.; Usón, I.; Laettig, S.; Welfle, H.; Sheldrick, G.M.; Saenger, W.; *Carbohydrate Research*. **2003**. *338(9)*. 977-986. (d) Gessler, K.; Usón, I.; Takaha, T.; Krauss, N.; Smith, S.M.; Okada, S.; Sheldrick, G.M.; Saenger, W.; *Proc. Natl. Acad. Sci. USA*. **1999**. *96(8)*. 4246-4251.
[127] (a) Gattuso, G.; Nepogodiev, S.A.; Stoddart, J.F.; *Chem. Rev*. **1998**. *98(5)*. 1919-1958. (b) Szejtli, J.; *Chem. Rev*. **1998**. *98(5)*. 1743-1754. (c) Biwer, A.; Antranikian, G.; Heinzle, E.; *Appl. Microbiol.*

As said, the process involves treatment of ordinary starch with a set of easily available enzymes[128] including, on the one hand, *Cyclodextrin Glycosyl Transferase*[129] (CGTase) which is present in many bacterial species, in particular, the *Bacillus genus* (e.g. *Bacillus Circulans*[130], *Bacillus Macerans*[131] and *Bacillus Stearothermophilus*), as well as some *Archaea*; and on the other hand, *α-amylase* (one of the *Glycoside Hydrolase Enzymes*) [Fig: 3.5].

The synthesis is carried out in two steps: in the first one, starch is liquefied either by heat treatment or using *α-amylase*, and then, in the second one, *CGTase* is added for the enzymatic conversion.



(a) **(b)**

Figure 3.5: Structures of (a) Cyclodextrin Glycosyl Transferase (CGTase) and (b) Human Salivary α-amylase[132].

Biotechnol. **2002**. *59(6)*. 609-617. (d) Zheng, M.; Endo, T.; Zimmermann, W.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*. 387-390. (e) Endo, T.; Zheng, M.; Zimmermann, W.; *Australian Journal of Chemistry*. **2002**. *55(2)*. 39-48.

[128] (a) Biwer, A.; Antranikian, G.; Heinzle, E.; *Appl Microbiol. Biotechnol*. **2002**. *59(6)*. 609-617. (b) van der Veen, B.A.; Uitdehaag, J.C.; Penninga, D.; van Alebeek, G.J.; Smith, L.M.; Dijkstra, B.W.; Dijkhuizen, L.; *J. Mol. Biol*. **2000**. *296(4)*. 1027-1038.

[129] (a) Uitdehaag, J.C.M.; van der Veen, B.A.; Dijkhuizen, L.; Dijkstra, B.W.; *Enzyme and Microbial Technology*. **2002**. *30(3)*. 295-304. (b) Kobayashi, S.; *Cyclodextrin Producing Enzyme (CGTase)*, in Park, K.H.; Robyt, J.F.; Choi, Y.-D.; (Eds.). *Enzymes for Carbohydrate Engineering*. Elsevier Science. Amsterdam. **1996**. 23–41. (c) Cucolo, G.R.; Alves-Prado, H-.F.; Gomes, E.; da Silva, R.; *Braz. J. Food Technol*. **2006**. *9(3)*. 201-208.

[130] Szerman, N.; Schroh, I.; Rossi, A.L.; Rosso, A.M.; Krymkiewicz, N.; Ferrarotti, S.A.; *Bioresource Technology*. **2007**. *98(15)*. 2886-2891.

[131] Tilden, E.B.; Adams, M.; Hudson, C.S.; *J. Am. Chem. Soc*. **1942**. *64(6)*. 1432-1433.

[132] Ramasubbu, N.; Paloth, V.; Luo, Y.; Brayer, G.D.; Levine, M.J.; *Acta Crystallogr. Sect. D*. **1996**. *52(3)*. 435-446.

One of the advantages of the enzymatic production is the high specificity: it is possible to improve the yield of a specific cyclodextrin by selecting the type of enzyme[133] because each *CGTase* has its own characteristic *α:β:γ* synthetic ratio, but also modifying other parameters like the incubation time, and the thermodynamic conditions (pH, temperature[134], and concentration). This is a suitable procedure not only for the synthesis of common cyclodextrins but also for the large ones, although the higher yield are unfortunately biased towards α-, β-, and γ-rings.

The final step, the purification process[127,135], is based on differences in water solubility: the value for β-CD is quite low (18.5 g/L at 25 ºC) in comparison with α- and γ-CDs ones (142 and 232 g/L at 25 ºC respectively), therefore, β-CD can be easily separated through crystallization while the more soluble α- and γ-CDs remain in solution waiting for a subsequent process, usually involving expensive and time consuming chromatographic and electrophoretic techniques[136].

It is also possible to improve purification yields by adding a complexing agent during the conversion process. Organic solvents[137] such as toluene, acetone or ethanol are commonly used to form an insoluble complex with the specific cyclodextrin. Then, equilibrium is displaced towards the synthesis of the precipitated macrocycle, which is separated by physical means (filtration or centrifugation). Afterwards, the complexing agent is chemically removed yielding the pure cyclodextrin.

3.5 Uses and Applications

There are different applications in a wide range of fields for the cyclodextrins and their derivatives[138], and almost all of them share a common ability to form host-guest complexes[139] given the amphiphylic nature imposed by their structure[140]:

[133] Terada, Y.; Sanbe, H.; Takaha, T.; Kitahata, S.; Koizumi. K.; Okada, S.; *Appl. Envirom. Microbiol*. **2001**. *67(4)*. 1453-1460.
[134] Qi, Q.; She, X.; Endo, T.; Zimmermann, W.; *Tetrahedron*. **2004**. *60(3)*. 799-806.
[135] (a) Ueda, H.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*. 53-56. (b) Ueda, H.; Wakisaka, M.; Nagase, H.; Takaha, T.; Okada, S.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*. 403-405.
[136] Larsen, K.L., Mathiesen, F.; Zimmermann, W.; *Carbohydrate Research*. **1997**. *298(1-2)*. 59-63.
[137] Qi, Q.; Mokhtar, M.N.; Zimmermann, W.; *J. Incl. Phenom. Macrocycl. Chem*. **2007**. *57(1-4)*. 95-99.
[138] (a) Del Valle, E.M.M.; *Process Biochemistry*. **2004**. *39(9)*. 1033-1046. (b) Hedges, A.R.; *Chem. Rev*. **1998**. *98(5)*. 2035-2044.

- Stabilisation of light- or oxygen-sensitive substances.
- Modification of the chemical reactivity of guest molecules.
- Fixation of very volatile substances.
- Improvement of solubility of substances.
- Modification of liquid substances to powders.
- Protection against degradation of substances by microorganisms.
- Masking of ill smell and taste.
- Masking pigments or the colour of substances.
- Catalytic activity of cyclodextrins with guest molecules.

The best known[141] pharmaceutical applications, and also probably the first ones in the industry, include those in which CDs are used as carriers for drug release [Fig. 3.6]: It is the case, for example, of *Johnson & Johnson's* oral antifungal ***Sporanox***® -containing itraconazol[142]-, *CINFA's* oral anticough ***Cinfatós Pastillas***®, -where β-cyclodextrin stabilizes the active drug dextrometorphan hydrobromide[143] and help in controlled releasing-, and *Schwarz Pharma's* impotence drug ***Edex***®[144], -where α-cyclodextrin solubilizes the E_1(PGE_1) prostaglandin; alprostadil[145]-. Other applications are those involving enantioselective chemical sensors[146] -able to distinguish between two enantiomers of the same molecule-. The industry of cosmetics is also interested in these molecules because of their already mentioned trend to encapsulate chemical compounds[138]. In general, solubilising[147] any kind of substrate.

[139] (a) Rekharsky, M.V.; Inoue, Y.; *Chem. Rev*. **1998**. *98(5)*. 1875-1917. (b) Larsen, K.L.; Endo, T.; Ueda, H.; Zimmermann, W.; *Carbohydrate Research*. **1998**. *309(2)*. 153-159.
[140] Harata, K.; *Chem. Rev*. **1998**. *98(5)*. 1803-1827.
[141] (a) Loftsson, T.; Duchêne, D.; *International Journal of Pharmaceutics*. **2007**. *329(1-2)*. 1-11. (b) Uekama,K.; Hirayama, F.; Irie, T.; *Chem. Rev*. **1998**. *98(5)*. 2045-2076. (b) Hid Cadena, R.; *inFÁRMAte*. **2008**. *Año 3. Número 17*. http://www.infarmate.org
[142] (a) **IUPAC**: (±)-1-[(*R**)-*sec*-butyl]-4-[*p*-[4-[*p*-[[(2*R**,4*S**)-2-(2,4-dichlorophenyl)-2-(1*H*-1,2,4-triazol-1-ylmethyl)-1,3-dioxolan-4-yl]methoxy]phenyl]-1-piperazinyl]phenyl]-Δ2-1,2,4-triazolin-5-one.
(b) Gregorí-Valdés, B.S.; *Rev. Cubana Farm*. **2005**. *39(2)*.
[143] **IUPAC**: (9alpha, 13alpha, 14alpha)-3-Methoxy-17-methylmorphinan hydrobromide.
[144] McCoy, M.; *C&EN*. **1999**. *77(9)*. 25-27.
[145] **IUPAC**: (1*R*,2*R*,3*R*)-3-Hydroxy-2-[(*E*)-(3*S*)-3-hydroxy-1-octenyl]-5-oxocyclopentane heptanoic acid.
[146] Shahgaldian, P.; Pieles, U.; *Sensors*. **2006**. *6(6)*. 593-615.
[147] Furuishi, T.; Endo, T.; Nagase, H.; Ueda, H.; Nagai, T.; *Chem. Pharm. Bull*. **1998**. *46(10)*. 1658-1659.

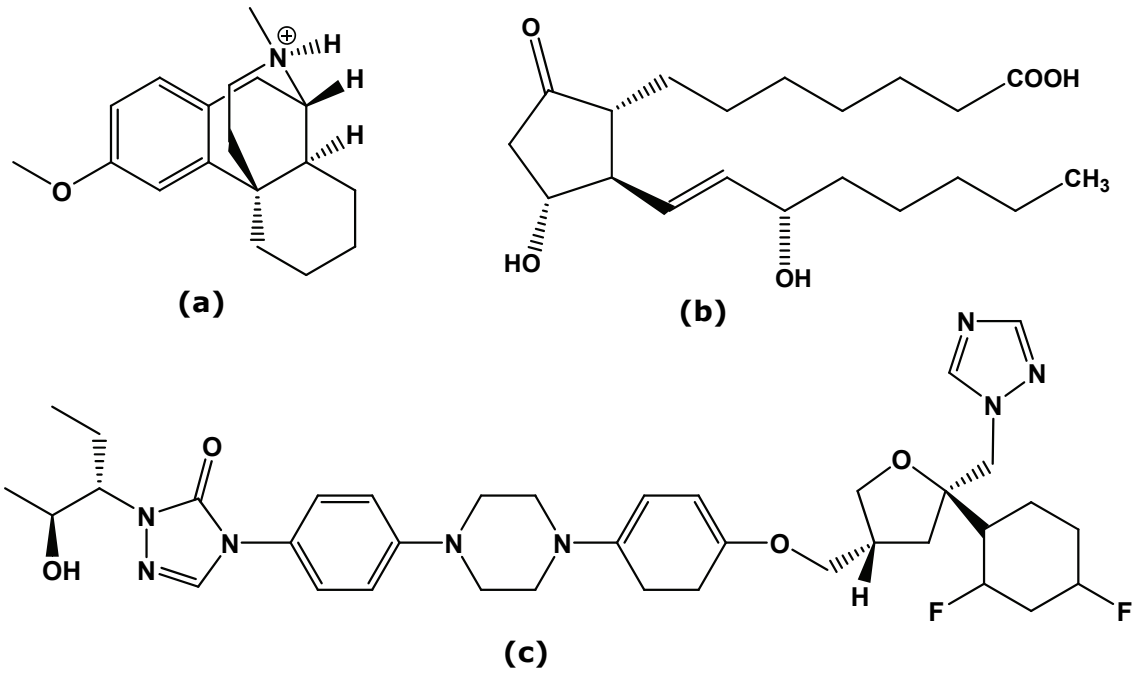


Figure 3.6: Examples of three drugs using cyclodextrins as vehicle: (a) Dextrometorphan hydrobromide, (b) Alprostadil and (c) Itraconazol.

In recent years, extending their growing interest to ecology and Nature, CDs have shown to be useful when employed in agriculture and environmental protection, immobilising toxic compounds[148] –organics, specially insecticides, pesticides, and heavy metals- inside their rings, or allowing effective complexation with stable substances, and acting as catalysts[149] enhancing further decomposition.

Besides, the ability to form stable complexes with cholesterol[150] made CDs to be easily introduced in the food industry for preparing cholesterol-free products[144]. Within this area, other applications have been discovered, including their ability to capture volatile or unstable compounds responsible for some unwanted tastes and odours [Fig. 3.7].

[148] (a) Sevillano Vaca, X.; *Doctoral Thesis*. Department of Chemistry and Edafology. Universidad de Navarra. **2006**. (b) Sevillano, X.; Romo, A.; Isasi, J.R.; Gonzalez-Gaitano, G.; Peñas, J.; *Waste Management and the Environment*. **2002**. 709-715. (c) Romo, A.; Peñas, F.J.; Sevillano, X.; Isasi, J.R.; *J. Ap. Polym. Sci*. **2006**. *100(4)*. 3393-3402. (d) Sevillano, X.; Isasi, J.R.; Peñas, F.J.; *Biodegradation*. **2008**. *19(4)*. 589-597. (e) Romo, A.; Peñas, F.J.; Isasi, J.R.; García-Zubiri, I.X.; González-Gaitano, G.; *React. Funct. Polym*. **2008**. *68(1)*. 406-413.

[149] Takahashi, K.; *Chem. Rev*. **1998**. *98(5)*. 2013-2033.

[150] (a) Rodal, S.K.; Skretting, G.; Garred, Ø.; Vilhardt, F.; van Deurs, B.; Sandvig, K.; *Molecular Biology of the Cell*. **1999**. *10(4)*. 961-974. (b) Abulrob, A.; Tauskela, J.S.; Mealing, G.; Brunette, E.; Faid, K.; Stanimirovic, D.; *Journal of Neurochemistry*. **2005**. *92(6)*. 1477-1486.

Figure 3.7: Example of a cyclodextrin capturing other substances such as odours or contaminants.

Soon, the point regarding odours opened the door to the group of industries in the field of fragrances and deodorants. On the one hand, CDs are employed as a carrier: pure solid cyclodextrin microparticles are exposed to active compounds, then, they are added to a variety of products such as fabrics, paints[151] or paper among many others. These devices are capable of releasing fragrances into the clothing -during ironing or when heated by human body- or inside buildings and houses –from the paints in the walls-. On the other hand, CDs are employed as a deodorant encapsulating odours: it is the case of *Procter and Gamble's* deodorizing product ***Febreze***©, where the main active compounds are CDs[144].

Coming back to the area of pure chemistry, several *"high-tech"* applications have been described. A few examples of them, in the field of supramolecular chemistry, were previously shown in the Introduction as an attempt to obtain molecular switches and similar devices when explaining certain *mechanically-interlocked molecular architectures*, such as *rotaxanes*[43,152] and *catenanes*[42]. In the area of synthesis, single cyclodextrins (and also dimers, tetramers…) and those carrying catalytic or reactive groups[153] have also been widely employed in biphasic catalysis[154], taking advantage of the differential polarity between the inner part –cavity- and the outer surface. Furthermore, several CDs have also been successfully applied in biphasic aqueous organometallic catalysis[155].

[151] Reisch, M.S.; *Chemical & Engineering News*. **2006**. *84(42)*. Web Exclusive.
[152] Nepogodiev, S.A.; Stoddart, J.F.; *Chem. Rev*. **1998**. *98(5)*. 1959-1976.
[153] Breslow, R.; Dong, S.D.; *Chem. Rev*. **1998**. *98(5)*. 1997-2011.
[154] Leclercq, L.; Sauthier, M.; Castanet, Y.; Mortreux, A.; Bricout, H.; Monflier, E.; *Adv. Synth. Catal*. **2005**. *347(1)*. 55-59.
[155] Leclercq, L.; Bricout, H.; Tilloy, S.; Monflier, E.; *J. Colloid Interface Sci*. **2007**. *307(2)*. 481-487.

In summary, the reason for this short review is to stress the growing importance of cyclodextrins in the industry, especially along the 80's and 90's. In words of József Szejtli, probably the most recognized authority in this field: *While in 1970 the price of 1 kg of β-CD was around $2000 US, and it was available only as a rare fine chemical, 25 years later, worldwide more than half a dozen companies are producing cyclodextrins. Their total output is in excess of 1000 tons/year, and the price of the key product, β-CD is only several dollars per kilogram, depending on quality and delivered quantity*[127].

The assertion can be easily proved just examining the literature in the field: The classical issue *Chem. Rev*. **1998** *.98(5)*, which is a monograph in cyclodextrins, was a quite well known *state-of-the-art* reference for any researcher in the area in that time. And although the last years have apparently witnessed a diminished interest in CDs, some other monographs such *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*; **2006**. *56(1-2)*; and **2007**. *57(1-4)*; and Helena Dodziuk's reviews book[156] have been published in the current decade reminding us that there are still plenty of possibilities for these molecules.

3.6 Molecular Structure. Glucose: The Basic Brick

As previously said, cyclodextrins are cyclic oligosaccharides composed of 5 or more α-D-glucopyranoside units linked 1→4, being the monosaccharide acting as basic unit the heterocyclic ring of glucose[157] [Fig. 3.8].

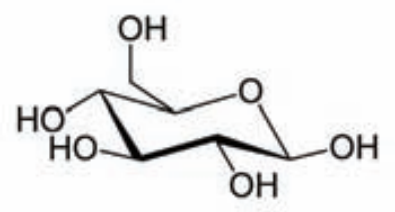


Figure 3.8: Glucose (Glc), is an important carbohydrate in biochemistry. Its name comes from the Ancient Greek word *glykys* (γλυκύς), meaning *sweet*, plus the suffix "-ose" which denotes *sugar*.

[156] *Cyclodextrins and Their Complexes*. Edited by Dodziuk, H.; WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. **2006**. ISBN: 3-527-31280-3.
[157] **IUPAC**: (2*R*,3*R*,4*S*,5*R*,6*R*)-6-(hydroxymethyl)tetrahydro-2*H*-pyran-2,3,4,5-tetraol.

3.6.1 Glucose

Glucose contains six carbon atoms, where the first one is part of an aldehyde group which makes the sugar to be referred as an *aldohexose*. In solution, the glucose molecule can exist in an open-chain *"acyclic"* form and in a ring *"cyclic"* form both in equilibrium, although the cyclic form is predominant at pH = 7 [Fig. 3.9].

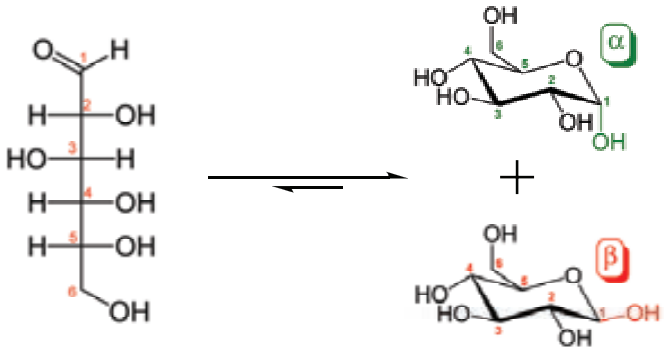


Figure 3.9: The cyclic form is the result of a covalent bond between the aldehyde C atom and the C-5 hydroxyl group to form a six-membered cyclic hemiacetal.

Because the ring contains five carbon atoms and one oxygen atom, which resembles the structure of pyran, the cyclic form of glucose is also referred to as *glucopyranose*. In this ring, each carbon is linked to a hydroxyl side group with the exception of the fifth atom, which links to a sixth carbon atom outside the ring, forming a -CH_2OH group [Fig.3.10].

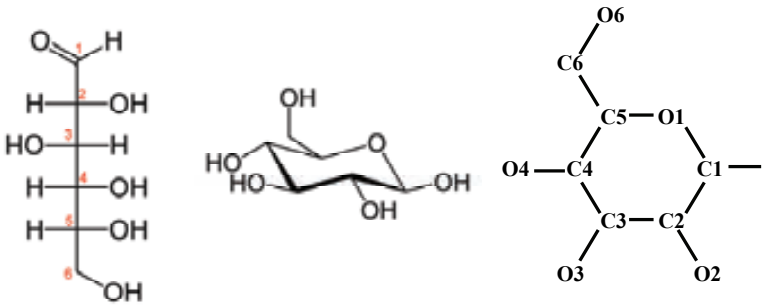


Figure 3.10: Lineal and cyclic molecule.

3.6.2 Isomers within the Aldohexose Family

All the aldohexose sugars are stereoisomers, which means that they share the same connectivity, molecular formula and molar mass, but they differ in the spatial arrangement of their groups around one or more sp^3 carbon atoms.

All in all, 4 chiral centers make up a total number of $2^4 = 16$ stereoisomers split into two groups, L and D, with 8 sugars in each. Only 8 out of 16 –the ones that belong to the D series- are found in living organisms, of which D-glucose (Glu), D-galactose (Gal), and D-mannose (Man) are the most important [Fig. 3.11]. These eight isomers -including glucose itself- are related as *diastereoisomers* and only D-(+)-glucose is the native structure and the biologically active one.

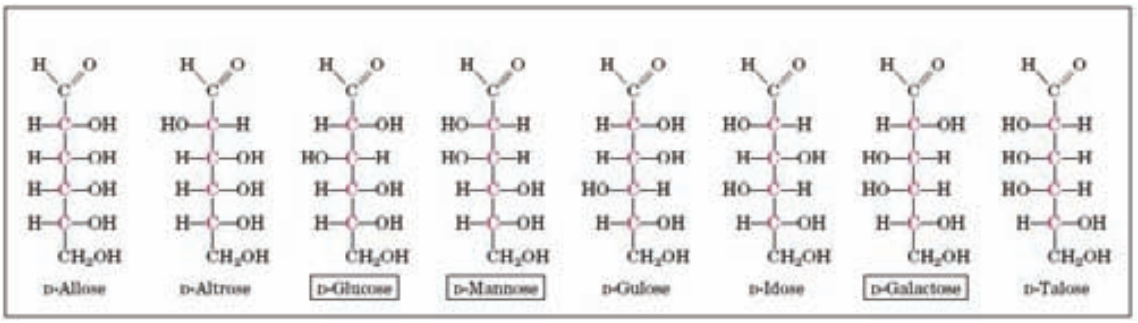


Figure 3.11: D-series of the Aldohexose family.

Furthermore, an additional asymmetric centre at C-1 -called the *anomeric carbon atom*- is created when glucose cyclizes and two ring structures called *anomers* are formed: α-glucose and β-glucose[158]. These anomers differ structurally by the relative positioning of the hydroxyl group linked to C-1, and the group at C-6 which is termed the reference carbon. The α and β forms interconvert over a timescale of hours in aqueous solution, to a final stable ratio of *α:β* 36:64, in a process called *mutarotation*[159] [Fig. 3.12].

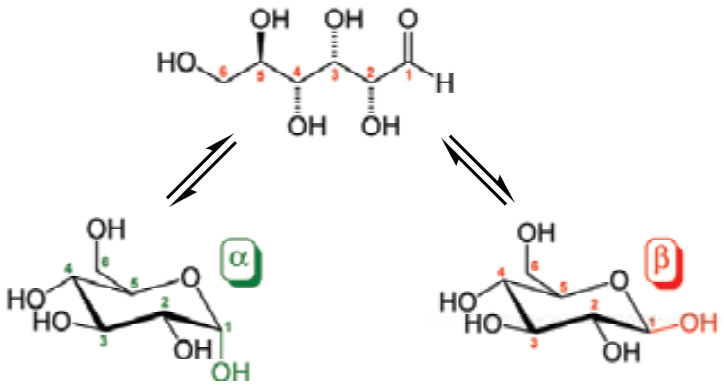


Figure 3.12: When D-glucose is drawn as a Haworth projection or in the standard chair conformation, the designation α means that the hydroxyl group attached to C-1 is positioned *trans* to the -CH_2OH group at C-5, while β means it is *cis*.

[158] Lee, Y.C.; Lee, R.T.; *J. Chin. Chem. Soc*. **1999**. *46(3)*. 283-292.
[159] Lehninger, A.L.; Nelson, D.L.; Cox, M.M.; *Principles of Biochemistry*. 4th Ed. W.H. Freeman & Company. **2004**. ISBN 9780716743392.

Although this work is exclusively based on macrorings entirely made up of glucoses, the point of this survey in sugars is to state the existence of macrocycles in the style of cyclodextrins based on monosaccharides[160] and monomers[161] other than glucose[116], i.e. the cyclomannoses[162], which are built up of mannose units, and also a variety of new artificial macrorings.

3.6.3 Rotamers of the D-(+)-glucose

This point relates the different behaviour regarding primary and secondary hydroxyl groups, although we will mainly focus on the secondary ones due to their higher flexibility.

Within the cyclic form of glucose, rotation may occur around the O6-C6-C5-O5 torsion angle, termed the ω-angle, to form three rotamer conformations as shown in the diagram below. For α-D-glucopyranose at equilibrium, the ratio of molecules in each rotamer conformation is reported as 57:38:5 gg:gt:tg[163]. This tendency for the ω-angle to prefer to adopt a gauche conformation is attributed to the *gauche effect*[164] [Fig. 3.13].

Figure 3.13: Referring to the orientations of the ω-angle and the O6-C6-C5-C4 angle the three stable staggered rotamer conformations are termed *gauche-gauche* (gg), *gauche-trans* (gt) and *trans-gauche* (tg).

[160] (a) Ashton, P.R.; Cantrill, S.J.; Gattuso, G.; Menzer, S.; Nepogodiev, S.A.; Shipway, A.N.; Stoddart, J.F.; Williams, D.J.; *Chem. Eur. J.* **1997**. *3(8)*. 1299-1314. (b) Ashton, P.R.; Brown, C.L.; Menzer, S.; Nepogodiev, S.A.; Stoddart, J.F.; Williams, D.J.; *Chem. Eur. J.* **1996**. *2(5)*. 580-591. (c) Chakraborty, T.K.; Srinivasu, P.; Bikshapathy, E.; Nagaraj, R.; Vairamani, M.; Kumar, S.K.; Kunwar*, A.C.; *J. Org. Chem*. **2003**. *68(16)*. 6257-6263.
[161] Fan, L.; Hindsgaul, O.; *Org. Lett*. **2002**. *4(25)*. 4503-4506.
[162] (a) Hid Cadena, R.; *inFÁRMAte*. **2008**. *Año 3. Número 17*. http://www.infarmate.org (b) Fukudome, M.; Shiratani, T.; Nogami, Y.; Yuan, D-.O.; Fujita, K.; *Org. Lett*. **2006**. *8(25)*. 5733-5736.
[163] (a) Kirschner, K.N.; Woods, R.J.; *Proc. Natl. Acad. Sci. USA*. **2001**. *98(19)*. 10541-10545. (b) Bock, K.; Duus, J.O.; *J. Carbohydr. Chem*. **1994**. *13(4)*. 513-543. (a) Nishida, Y.; Ohrui, H.; Meguro, H.; *Tetrahedron Lett*. **1984**. *25(15)*. 1575-1578.
[164] Craig, C.; Chen, A.; Suh, K.H.; Klee, S.; Mellau, G.C.; Winnewisser, B.P.; Winnewisser, M.; *J. Am. Chem. Soc*. **1997**. *119(20)*. 4789-4790.

Not many things are going to be explained here about hydrogen bonding and rotamers, but this point will be of importance in the next chapters when talking about native CDs, large CDs and folding.

3.7 Nomenclature and Geometrical Parameters

The *Joint Commission on Biochemical Nomenclature* has proposed several rules for the standard nomenclature of polysaccharides and geometrical parameters such as atom numbers, distances, angles, and dihedrals[165].

3.7.1 Lineal Chains

Regarding lineal chains, the conventional depiction has the reducing sugar -glycose residue- on the right (the first residue: 1) and the non-reducing end -glycosyl group- on the left (the last residue: n), being the internal sugar units called glycosyl residues (2, 3, 4… n-1) [Fig. 3.14].

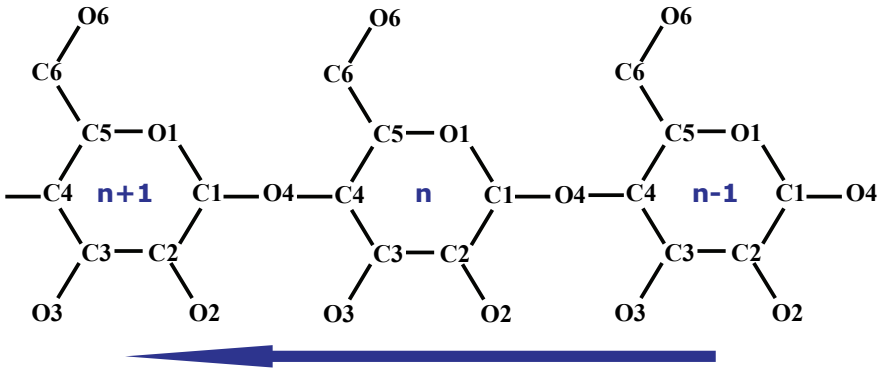


Figure 3.14: Numbering of residues in a lineal chain.

3.7.2 Cyclodextrins

When referring to cyclodextrins, there is no "first" or "last residue", just because the macromolecule is a ring and all the residues within it possess the same priority. In this

[165] (a) IUPAC. *Pure & Appl. Chem*. **1983**. *55(8)*. 1269-1272. (b) IUPAC-IUB (JCBN). *Eur. J. Biochem*. **1983**. *131(1)*. 5-7. (c) IUPAC. *Pure & Appl. Chem*. **1996**. *68(10)*. 1919-2008. (d) McNaught, A.D.; *Carbohydr. Res*. **1997**. *297(1)*. 1-92.

situation, the “direction” of the naming is maintained as in the case of equivalent opened-chain polysaccharides and the counting goes from the first residue -beginning arbitrarily at any of the glucoses- until the same unit is reached at the other side [Fig. 3.15].

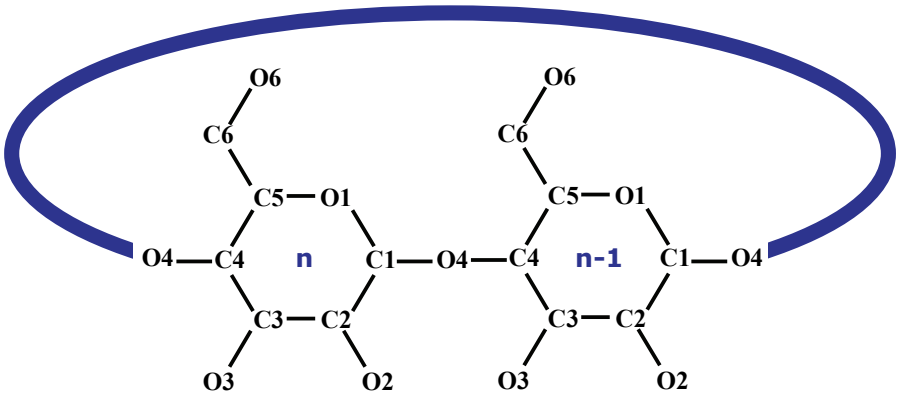
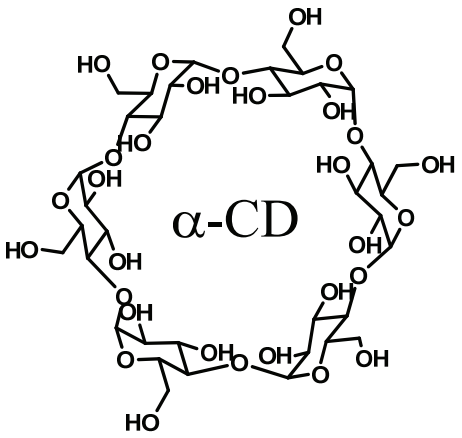


Figure 3.15: Numbering of residues in a cyclic chain.

Currently, there are almost no doubts in the nomenclature of cyclodextrins since IUPAC clearly established the semisystematic and systematic names for cyclic polysaccharides in *Pure & Appl. Chem.* **1996**. *68(10)*. 1919-2008. already cited in reference 165(c):

3.7.2.1 Semisystematic Names

Cyclic oligosaccharides composed of a single type of oligosaccharide unit may be named semisystematically by citing the prefix “cyclo”, followed by terms indicating the type of linkage [e.g. “malto” for α-(1→4)-linked glucose units], the number of units (e.g. “hexa” for six) and the termination “-ose”. The trivial names α-cyclodextrin (α-CD) for cyclomaltohexaose, β-cyclodextrin (β-CD) for cyclomaltoheptaose and γ-cyclodextrin (γ-CD) for cyclomaltooctaose are well established [Fig. 3.16].

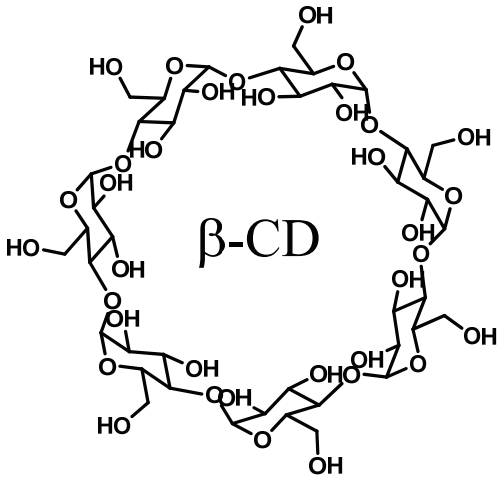


Cyclomaltohexaose
(α-cyclodextrin, α-CD)

Figure 3.16: Semisystematic name for the α-cyclodextrin.

3.7.2.2 Systematic Names

Cyclic oligosaccharides composed of a single type of residue can be named by giving the systematic name of the glycosyl residue, preceded by the linkage type in parentheses, preceded in turn by "cyclo-" with a multiplicative suffix (i.e. "cyclohexakis-" etc.) [Fig. 3.17].



Cycloheptakis-(1→4)-α-D-glucosyl

Figure 3.17: Systematic name for the β-cyclodextrin.

3.7.3 Names and Geometrical Parameters

As seen, IUPAC nomenclature takes into account the cyclic nature of cyclodextrins, the type of bonding and the number of glucose units. This is useful for describing the sort of macromolecule we are working with, but quite often these names are neither easy to

handle nor they provide any conformational information -which is really important when describing 3D properties-. In this sense, IUPAC did half the work when proposing the standard numbering order in the atoms within the glucose units, and the different geometrical parameters –angles, dihedral angles, and distances- to be considered as standards for conformational studies. The rest of the work is our job:

Since this Ph.D. Thesis is a conformational study, and long names are not usually easy handling, 2 rules will be employed hereafter on the basis of *simplicity* and *compatibility* to designate the molecules unambiguously:

3.7.3.1 Simplicity

Cyclodextrins in the present work will usually be referred as "CDn", using a standard widely-accepted non-systematic nomenclature[166], where CD is the abbreviation of "cyclodextrin", and "n" stands for the total number of glucose units within the macroring [Fig. 3.18].

Glycosyl	Names			Abbreviations		
Units	IUPAC Semisystematic	IUPAC Systematic	Generic	*greek*	*CA*	*CD*
5	cyclomaltopentaose	Cyclopentakis-(1→4)-α-D-glucosyl	-	-	CA5	CD5
6	cyclomaltohexaose	Cyclohexakis-(1→4)-α-D-glucosyl	α-cyclodextrin	α-CD	CA6	CD6
7	cyclomaltoheptaose	Cycloheptakis-(1→4)-α-D-glucosyl	β-cyclodextrin	β-CD	CA7	CD7
8	cyclomaltooctaose	Cyclooctakis-(1→4)-α-D-glucosyl	γ-cyclodextrin	γ-CD	CA8	CD8
14	cyclomaltotetradecaose	Cyclotetradecakis-(1→4)-α-D-glucosyl	ι-cyclodextrin	ι-CD	CA14	CD14
21	cyclomaltoheneicosaose	Cycloheneicosakis-(1→4)-α-D-glucosyl	π-cyclodextrin	π-CD	CA21	CD21
26	cyclomaltohexaicosaose	Cyclohexaicosakis-(1→4)-α-D-glucosyl	φ-cyclodextrin	φ-CD	CA26	CD26
28	cyclomaltooctaicosaose	Cyclooctaicosakis-(1→4)-α-D-glucosyl	ψ-cyclodextrin	ψ-CD	CA28	CD28

Figure 3.18: Table containing different nomenclatures for the cyclodextrins employed in this thesis.

3.7.3.2 Compatibility

The Geometrical parameters studied in this work are directly based on those proposed by IUPAC[165(b)]. This ensures full parameter concordance when comparing present results with already published data [Fig. 3.19].

[166] Larsen, K.L.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *43(1-2)*. 1-13.

Distances	O3(n)	O2(n+1)			
	O4(n)	O4(n+1)			
Angles	C4(n)	O4(n)	C1(n+1)		
	O4(n)	O4(n+1)	O4(n+2)		
Dihedrals	C1(n)	O4(n-1)	C4(n-1)	C3(n-1)	
	O3(n)	C4(n)	C1(n+1)	O2(n+1)	**[flip]**
	O4(n)	O4(n+1)	O4(n+2)	O4(n+3)	**[Ψ]**
	O1(n)	C1(n)	O4(n-1)	C4(n-1)	**[Φ]**

Figure 3.19: Table of standard parameters.

3.8 Benchmark: Common and Large Cyclodextrins

3.8.1 The Set of Molecules

In summary, the family of macromolecules selected in the present work to test our conformational search methodology is that set of rings previously studied in our group by Dr. Ivan Beà and Dr. Itziar Maestre, plus the smaller 5-membered ring. The set can be divided in 2 groups according to their flexibility. On the one hand, the native CDs including those containing 6, 7, and 8 glucose units, and the smaller structure with 5 [Fig. 3.20].

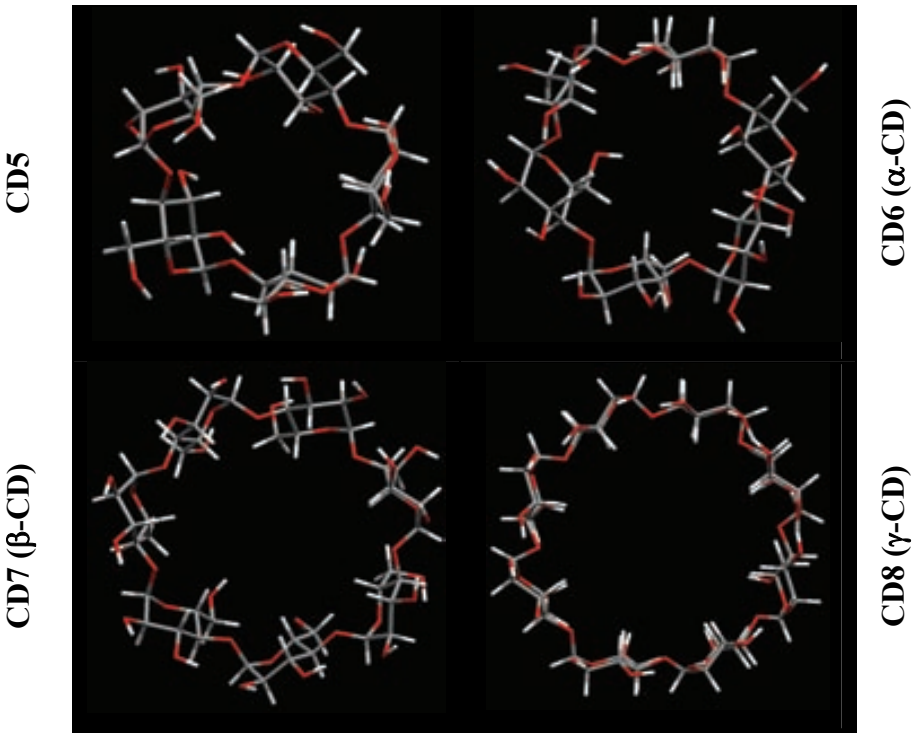


Figure 3.20: Table containing the small cyclodextrins; the 5-membered ring and the native ones.

On the other hand, the larger ones containing 14, 21, 26 and 28 glucose units [Fig. 3.21].

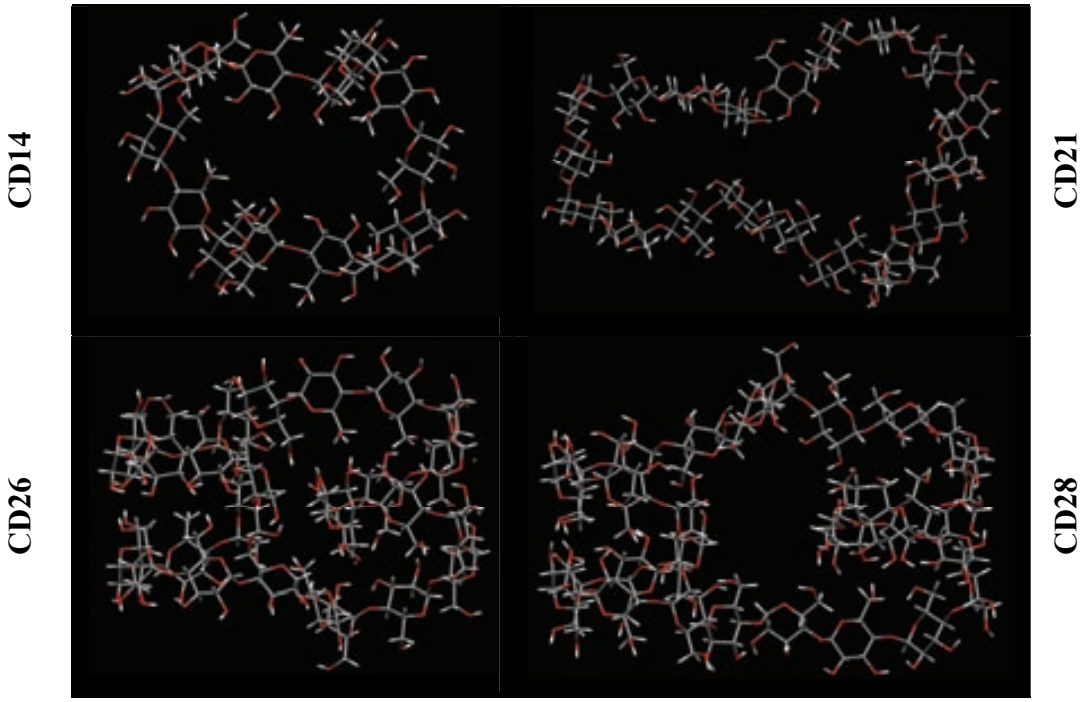


Figure 3.21: Table containing the large cyclodextrins.

3.8.2 Hydrogen Bonding and Secondary Structure

The previous classification is not random at all due to the different behaviour among the molecules in those groups. Significant conformational changes in the series have been reported when studying cyclodextrins over 9 glucose units[167], because those molecules exhibit *Secondary Structure* and therefore they may properly be considered as *Foldamers*.

The foldameric behaviour was already outlined in the Introduction and clearly divides the set in 2 groups: that one in which molecules can fold and that one in which molecules cannot.

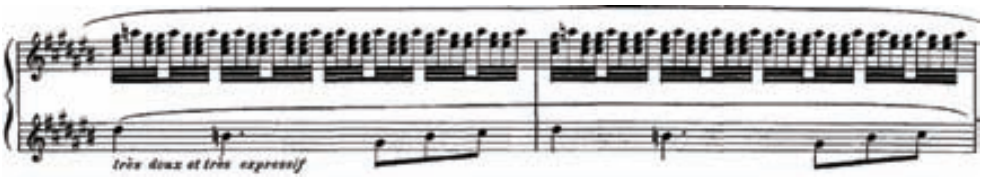
This feature is one of highly relevance when considering the flexibility and the total number of conformations as will be discussed in the forthcoming chapters (Results).

[167] Saenger, W.; Jacob, J.; Gessler, K.; Steiner, T.; Hoffmann, D.; Sanbe, H.; Koizumi, K.; Smith, S.M.; Takaha, T.; *Chem. Rev*. **1998**. *98(5)*. 1787-1802.

While conformational changes in the smaller molecules are closely related to the interactions of primary and secondary hydroxyl groups among glucoses in the nearest vicinity[168] ("n-1", "n", and "n+1" glucoses), the scenario is absolutely different regarding the larger ones, allowing effective Hydrogen bonding interactions involving any two glucoses farther in the primary structure. This new situation regards *rotamers*[169] as only a small part in the whole picture of conformational search.

[168] (a) Lesyng, B.; Saenger. W.; *Biochim. et Biophys. Acta*. **1981**. *678(3)*. 408-413. (b) Koehler, J.E.H.; Saenger, W.; Lesyng, G.; *J. Comput. Chem*. **1987**. *8(8)*. 1090-1098. (c) Saenger, W.; *Nature*. **1979**. *279(5711)*. 343-344.
[169] Jiménez, V.; Alderete, J.B.; *J. Phys. Chem. A*. **2008**. *112(4)*. 678-685.

RESULTS



Maurice Ravel
Gaspard de la Nuit
(Trois poèmes pour piano d'après Aloysius Bertrand)
(1909)
I. Ondine

"Your theory is crazy... but it's not crazy enough to be true."

Niels Bohr

4 RESULTS I: METHODOLOGY. NOMENCLATURE AND DATA MINING IN CYCLODEXTRINS

4.1 CLASSIFICATION: INTRODUCTION

At this point, let us remember the topics already discussed: the objectives were explained in Chapter I; the general Introduction was included in Chapter II, and finally, the molecules selected as benchmark –the family of cyclodextrins- were introduced in Chapter III.

Now, the present Chapter deals with the *In Home* Methodologies and Techniques developed for the analysis of those sets of conformations: the ***Nomenclature for Cyclodextrins*** and the ***Data Mining and Classification Tools***. A more general approach to the Methodological foundations of Molecular Modelling is included in Chapter IX as an appendix.

4.2 MOLECULAR DESCRIPTOR: NOMENCLATURE

When dealing with small molecules we usually refer to specific conformations and stereospatial arrangements using standard descriptors such as *Cahn-Ingold-Prelog* and *Z/E* indexes, or *syn-anti*, *gauche*-(+),(-), *eclipsed* terminology. Their combination with projections such as Newmann, Fischer, Isometric, 3D-Tetrahedral, Haworth, and Natta aid chemists in drawing chemical structures. In a sense, all of them can be regarded as "optical" indexes, because they lead to a 3D picture of the molecule in our mind.

Macromolecules exhibit a high flexibility that enables them to spread *almost* continuously over all their conformational spaces. This makes quite a difference in comparison with small molecules, which usually fall into well-established patterns of conformational groups. Therefore, those tools turn out to be useless on their own –or at least less specific, when they are used to depict conformational information- as the size of the molecules under study grows larger. As said before, the scale-factor is of great importance in conformational analysis.

So, the question is: is it possible to create any useful spatial representation of larger molecules in our mind using only those indexes? The answer is: No. We need another

way to gather the particular features of any molecule irrespective of their size and connectivity; an effective index or set of indexes: that is, the *Molecular Descriptors*[170].

4.2.1 Molecular Descriptors and 3D structure.

In words of Todeschini and Consonni[171], researchers in the field and creators of E-DRAGON[172] software: *"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment"*. In our particular case, it can also be defined as *any numerical or alphabetical index that uniquely contains -and explains- the larger amount of conformational information with the higher degree of compactness*.

These indexes are of great importance and there are many choices available -i.e. E-DRAGON provides more than 1,600 molecular descriptors divided into 20 logical blocks-.

Molecular Descriptors are commonly rather mathematical in comparison to those previously mentioned. The reason for this is to allow further treatment as they are often Data Mining tools[173] that enable searching through databases[174] –as in QSAR research- and also prediction under reverse-codifying data from the Molecular Descriptor Space[175].

In the present work, a new enhanced molecular descriptor for cyclodextrins is proposed as a tool for conformational analysis.

---

[170] Arteca, G.A.; *Molecular Shape Descriptors*; *Reviews in Computational Chemistry*. **1996**. *9.* Chapter. 5.
[171] Todeschini, R.; Consonni, V.; *Handbook of Molecular Descriptors*. Wiley-VCH. **2000**. ISBN 3-52-29913-0.
[172] (a) Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; Tanchuk, V.Y.; Prokopenko, V.V.; *J. Comput. Aid. Mol. Des*. **2005**. *19(6)*. 453-463. (b) http://www.vcclab.org/lab/edragon/
[173] Bajorath, J.; *J. Chem. Inf. Comput. Sci*. **2001**. *41(2)*. 233-245.
[174] Ros, F.; Pintore, M.; Chrétien, J.R.; *Chemometrics and Intelligent Laboratory Systems*. **2002**. *63(1)*. 15-26.
[175] Visco Jr., D.P.; Pophale, R.S.; Rintoul, M.D.; Faulon, J-.L.; *Journal of Molecular Graphics and Modelling*. **2002**. *20(6)*.429-438.

4.2.2 Descriptor I: Nomenclature by Dr. Itziar Maestre.

The first attempt to define an adequate conformational descriptor in our group was made by Dr. Itziar Maestre. The starting point can be summarized in the next assumption:

"Cyclodextrins conformational information –including folding- can be successfully explained just on the grounds of dihedral angles between the glucose units".

$$Conformation = f\left(array\left[dihedralAngles\right]\right)$$

Dr. Maestre chose the *flip* dihedral [Fig. 4.1] as the chief parameter and defined the array of flip dihedrals for a given CD as a raw Molecular Descriptor. The so defined index was a continuous n-dimensional array on the Real Numbers Field, $\mathbf{R}^{n}$, where "n" stands for the number of dihedral angles (that also equals the number of glucose units).

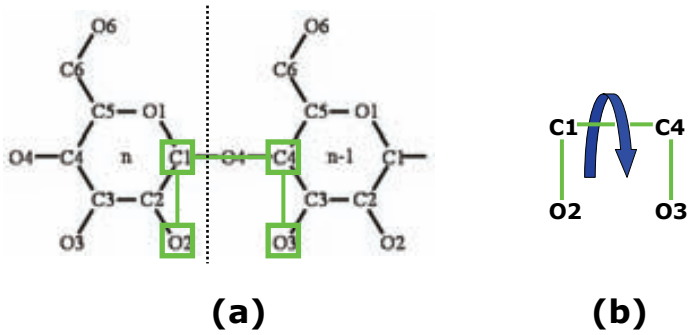


Figure 4.1: (a) *Flip* Dihedral Angle between glucoses as defined by IUPAC and starting point for Dr. Maestre's Descriptor. (b) Atoms involved in *flip* dihedral.

In the next step the components in the array were staggered into discrete classes to achieve a more simplified and easy-handling descriptor. Therefore, a "similar" nomenclature based on that one IUPAC proposed for dihedral angles was created [Fig. 4.2].

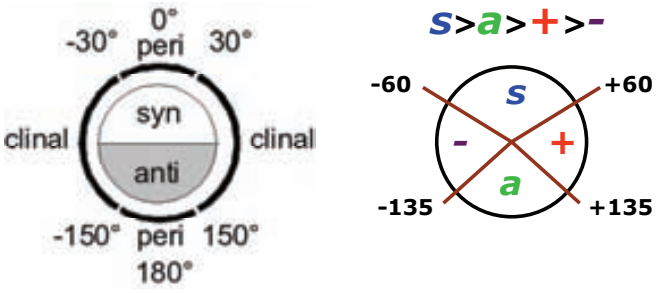


Figure 4.2: (a) Discrete nomenclature for dihedrals by IUPAC. (b) Discrete nomenclature for *Flip* Dihedrals by Dr. Maestre.

The last step was establishing the precedence rules for naming conformations according to the discrete characters and the sequence of dihedrals:

4.2.2.1 Rules of Precedence:

1) The priority is "arbitrarily" set to: **s** > **a** > + > -. Then, the name is built as a sequence of characters separated by *commas*. I.e. [**s,s,s,s,+,+,a,-**].
2) Sequences of identical characters are reduced to a single sequence where the character is preceded by a numerical index that accounts for the number of times the symbol is repeated: I.e. [**s,s,s,s,+,+,a,-**] is transformed into [**4s,2+,a,-**].
3) Since CDs are cyclic molecules, there is a rotational symmetry in the chain of characters that must be removed. Therefore, to avoid this problem, the descriptor *must* always start beginning with the largest [**s**]-sequence. I.e. the descriptor [**2s,+,-,3s,a**] should be rewritten in this way [**3s,a,2s,+,-**].
4) When several equivalent beginning sequences are possible -as in C.I.P indexes- the precedence is decided by the next sequence to the right, and so on, until the best sequence is determined. I.e. [**2s,-,+,2s,a,**+]. In this case, the sequence [**2s**] happens to occur twice. Then we continue next to the right and we find a [**-**] character after the first [**2s**] sequence, and [**a**] character after the second [**2s**] sequence. Since the first rule says that the precedence of [**a**] is higher than [**-**], the starting point of the name is the [**2s**] sequence followed by the [**a**] character. Then the correct name is [**2s,a,+,2s,-,+**].

   In this sense, the priority is firstly determined by the precedence of characters, and secondly by the indexes. I.e. In case that [**…2s,5-…**] and [**…2s,a…**] were

found, the second name would be *always* preferred as the starting point irrespective of the indexes because [**a**] has higher precedence than [**-**].

This descriptor –herein after called D-I- was used by Dr. Itziar Maestre to identify groups of conformations obtained from MD calculations.

4.2.3 Descriptor II: Nomenclature in this work.

Extensive work in Conformational Search and Molecular Dynamics Analysis was to be done in the present research, so we needed a trustworthy molecular descriptor that would fully and effectively describe the conformational features of our molecules. This descriptor would be used as a comparison tool and our first choice was, obviously, D-I.

The process of analysis at the beginning of our research relied on D-I but, shortly after, we detected one inconsistency. Detailed examination of flip dihedral histograms showed that the division of classes had been mistakenly done:

1) The borders were not coincident with the minima of the distributions and,
2) The main class was incorrectly defined, encompassing the two most important distributions within the same character, [**s**]; thus, different conformations would be considered the same [Fig. 4.3].

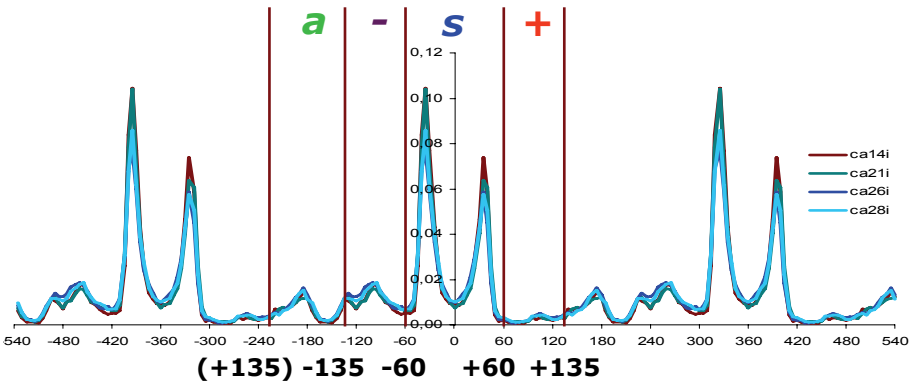


Figure 4.3: Superimposed Histograms of *Flip* Dihedral data obtained from large Cyclodextrins –data plotted 3 times-; division of classes by Dr. Maestre.

The small incongruence about the borders was a minor setback, but the problem regarding the [**s**]-class directly pointed to the very foundations of the D-I definition.

To solve this problem, we decided to create a new descriptor improving that one proposed by Dr. Itziar Maestre. Two measures were adopted:

1) The number of classes was enlarged to 5 [Fig. 4.4]: *We adapted ourselves to Nature, instead of forcing Nature to adapt to ourselves*.

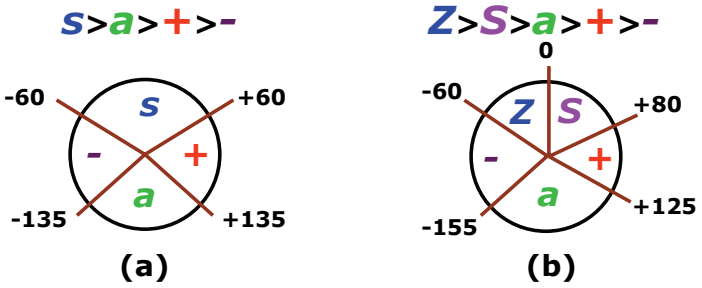


Figure 4.4: (a) Nomenclature for *Flip* Dihedrals by Dr. Maestre. (b) Improved nomenclature employed in this work.

2) The borders were shifted to fit the best the minima of the Histogram [Fig. 4.5].

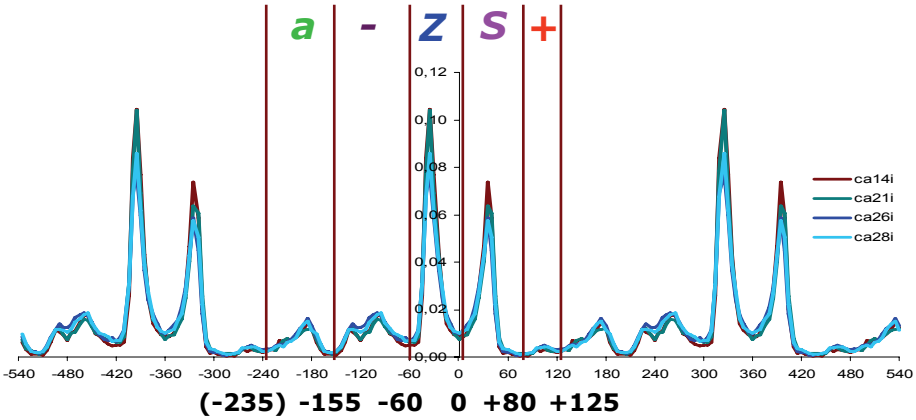


Figure 4.5: Superimposed Histograms of *Flip* Dihedral data obtained from large Cyclodextrins –data plotted 3 times-; new division of classes proposed in this work.

Therefore, the precedence rules for Descriptor II –herein after called D-II- were re-written, modifying those established for D-I:

4.2.3.1 Rules of Precedence:

1) The priority is "arbitrarily" set to: **Z > S > a > + > -**. Then, the name is built as a sequence of characters separated by *commas*. I.e. [**Z,Z,S,S,+,+,a,-**].

2) Sequences of identical characters are reduced to a single sequence where the character is preceded by a numerical index that accounts for the number of times the symbol is repeated: I.e. [**Z,Z,S,S,+,+,a,-**] is transformed into [**2Z,2S,2+,a,-**].
3) Since CDs are cyclic molecules, there is a rotational symmetry in the chain of characters that must be removed. Therefore, to avoid this problem, the descriptor *must* always start beginning with the largest [**Z**]-sequence. I.e. the descriptor [**2Z,+,-,3Z,S**] should be rewritten in this way [**3Z,S,2Z,+,-**].
4) When several equivalent beginning sequences are possible; the precedence is decided by the next sequence to the right, and so on, until the best sequence is determined. I.e. [**2Z,-,+,2Z,a,+**]. In this case, the sequence [**2Z**] happens to occur twice. Then we continue next to the right and we find a [**-**] character after the first [**2Z**] sequence, and [**a**] character after the second [**2Z**] sequence. Since the first rule says that the precedence of [**a**] is higher than [**-**], the starting point of the name is the [**2Z**] sequence followed by the [**a**] character. Then the correct name is [**2Z,a,+,2Z,-,+**].

   In this sense, the priority is firstly determined by the precedence of characters, and secondly by the indexes. I.e. In case that [**…2Z,5-…**] and [**…2Z,a…**] were found, the second name would be *always* preferred as the starting point irrespective of the indexes because [**a**] has higher precedence than [**-**].

In all cases within the present work, both D-I and D-II, have been automatically calculated using the software ***NameGiantCyD*** [Fig. 4.6]. The full code is included in the Appendix section and can be checked in Chapter 11.

D-II was our new and definitive choice but we wanted to be sure about it. The new definition was mathematically tested. In order to prove how reliable this index was, further Principal Component Analysis was performed to check its goodness-of-fit.

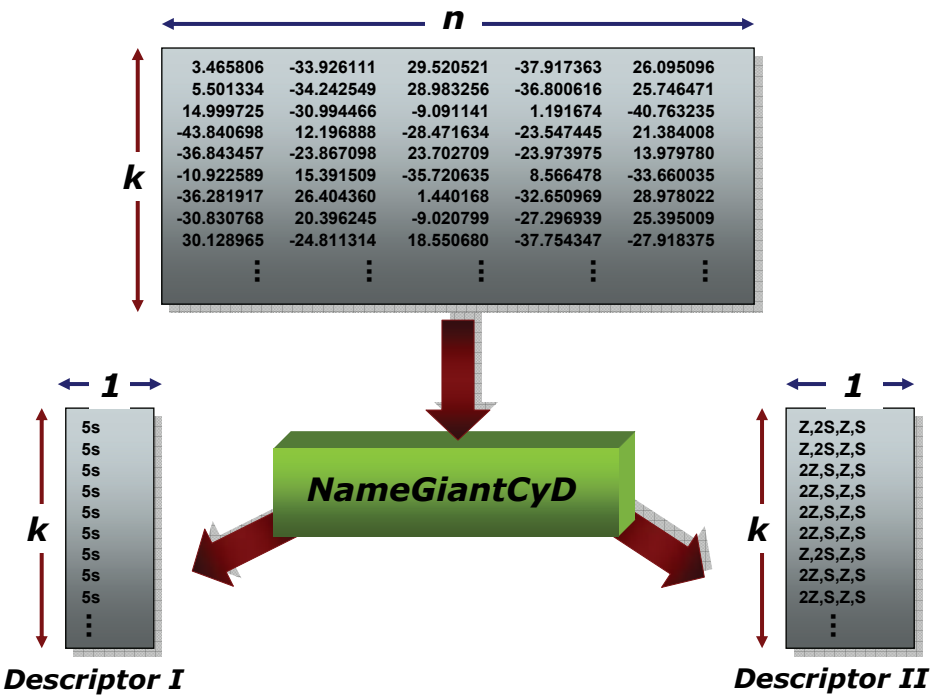


Figure 4.6: Flow chart of *NameGiantCyD*. The input file contains the *k* x *n* matrix of data, in which *k* is the number of conformations and *n* the number of glucoses. The program processes the file and produces a stream containing a *k* x *1* matrix –array- of descriptors by the standard output. It is possible to select the type of descriptor –D-I or D-II- by passing the appropriate flag on the command line.

4.2.4 Principal Component Analysis (PCA). Checking Descriptor II.

4.2.4.1 "Good enough is good".

Every time you map a continuous function onto a discrete space, the total amount of information is reduced, but the simplicity and generality you gain compensates it: *"Good enough is good"*.

When a discrete descriptor is created as in our case, dihedral values within the same distribution function are considered *similar*, or *fluctuations around a central value*, so they can be represented by a single character. This means a reduction of information but not really a lost of information. This process still keeps, in a sense, a ***bijective*** property: *a one-to-one correspondence between the set of conformations and the set of names*.

D-I did not accomplish that bijective correspondence and D-II was created to solve this drawback.

4.2.4.2 Principal Component Analysis.

PCA was firstly described by Pearson[176], and is a mathematical technique in the area of statistics that attempts to determine a smaller set of synthetic variables that could explain the original set[177]. It operates two changes in the original data:

1) Transform the original variables of the system into a new set of variables –the Principal Components- that show a *decoupled* behaviour, and
2) Reduce the dimensionality of the system.

Essentially, PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance[178] (rotation and scaling). The algorithm, briefly explained, involves several steps:

1) Calculation of the covariance matrix.
2) Diagonalisation of the covariance matrix.
3) Sorting eigenvalues and eigenvectors in decreasing order.

Since eigenvalues represent the percentage of variance explained and eigenvectors – principal components- show the relationship between the older variables and the new ones, detailed examination of this data allow better comprehension of the system.

4.2.4.3 Test of comparison: PCA vs. Descriptor II.

PCA is widely applied in studies involving Conformational Analysis in order to separate groups of conformations[179]; like in our case. This methodology has been introduced in our group by ***Felicidad Franch Lage***, who has personally done the

[176] Pearson, K.; *Phil. Mag.* **1901**. *2(6)*. 559-572.
[177] Shlens, J.; *"A Tutorial on Principal Component Analysis"*. **2005**. Systems Neurobiology Laboratory, Salk Insitute for Biological Studies. La Jolla, CA 92037 and Institute for Nonlinear Science, University of California, San Diego La Jolla, CA 92093-0402. http://www.snl.salk.edu/~shlens/notes.html.
[178] Ramis Ramos, G.; García Álvarez-Coque, Mª.-C.; *Quimiometría*. Editorial Síntesis. **2001**. ISBN 84-7738-904-7.
[179] (a) Glen, W.G.; Dunn III, W.J.; Scott, D.R.; *Tetrahedron Computer Methodology*. **1989**. *2(6)*. 349-376. (b) Glen, W.G.; Sarker, M.; Dunn III, W.J.; Scott, D.R.; *Tetrahedron Computer Methodology*. **1989**. *2(6)*. 377-396.

analysis of CD5 and CD6, and has collaborated helping us when we have done the others.

1) 1024 steps SA calculation for CA5 is going to be analysed.
2) Conformational results obtained by means of PCA and by means of D-II Saturation Analysis (see 5.3.2) will be compared.
3) In case that both the methodologies led to the same results, then D-II could be considered as correctly defined and nomenclature would be also suitable for larger CDs.

This methodology will be applied to the *flip* dihedral raw Molecular Descriptor using Multivariate Analysis Software Package **Unscrambler 9.x**[180]. As said, this n-array of Real Numbers contains all the conformational information but still there is the important problem regarding the cyclic symmetry of the cyclodextrins.

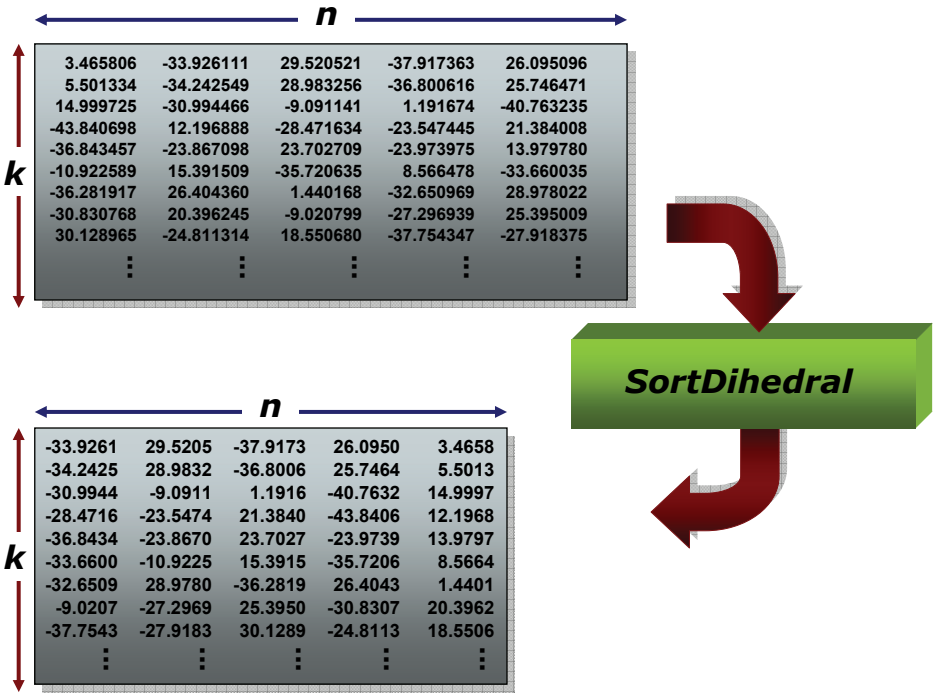


Figure 4.7: Flow chart of *SortDihedral*. The input file contains the *k* x *n* matrix of data, being *k* the number of conformations and *n* the number of glucoses. The program processes the file and produces a stream containing a *k* x *n* matrix, correctly sorted, by the standard output. It is possible to select the type of ordering by passing the appropriate flag on the command line.

Due to the fact that PCA cannot be directly applied to our data, a previous treatment -in order to compare all the arrays under the same conditions- is necessary. Basically, this

[180] *Unscrambler™*. **1986-2004**. *CAMO PROCESS AS*. http://www.camo.com.

treatment consists in removing the cyclic symmetry of every conformation – Cyclodextrins can be represented by the point group[181] C_n- by rotation of the array.

The sorting criteria are similar to the preference rules for Descriptors I & II –both of them are available-. However, in this case they are directly applied to dihedral values instead of sequences. All the data within the present work has been automatically processed using the software ***SortDihedral*** [Fig. 4.7]. The full code is also included in the Appendix section; Chapter 11.

The next example is shown here just to check the good-of-fitness of D-II.

4.2.4.4 CD5

Data from 1024 steps *long*-SA calculations were processed. CD5, the smaller cyclodextrin studied in this work, was selected for this test because of its higher rigidity and easy-handling conformational space.

The next table [Fig. 4.8] includes the headings and the last line of three files: the flip dihedral values, the name according to Descriptor I, and the name according to Descriptor II. Detailed examination of the list of names clearly show that under Descriptor I criteria, all of them are similar [**5s**], while under Descriptor II criteria two different conformations are found; [**Z,2S,Z,S**] and [**2Z,S,Z,S**].

Nº Conf.	"flip" Dihedrals from ca05iSA1025 (raw Data)					Descriptor I	Descriptor II
1	3.465806	-33.926111	29.520521	-37.917363	26.095096	5s	Z,2S,Z,S
2	5.501334	-34.242549	28.983256	-36.800616	25.746471	5s	Z,2S,Z,S
3	14.999725	-30.994466	-9.091141	1.191674	-40.763235	5s	2Z,S,Z,S
4	-43.840698	12.196888	-28.471634	-23.547445	21.384008	5s	2Z,S,Z,S
5	-36.843457	-23.867098	23.702709	-23.973975	13.979780	5s	2Z,S,Z,S
6	-10.922589	15.391509	-35.720635	8.566478	-33.660035	5s	2Z,S,Z,S
7	-36.281917	26.404360	1.440168	-32.650969	28.978022	5s	Z,2S,Z,S
8	-30.830768	20.396245	-9.020799	-27.296939	25.395009	5s	2Z,S,Z,S
9	30.128965	-24.811314	18.550680	-37.754347	-27.918375	5s	2Z,S,Z,S
...	...	...	...	...	...	...	...
1025	20.634097	-37.714777	-27.209981	28.482221	-25.733136	5s	2Z,S,Z,S

Figure 4.8: CD5 1024 steps Simulated Annealing calculation.
Conformations similar under D-I criteria are different under D-II criteria.

[181] Cotton, F.A.; *Chemical Applications of Group Theory*. 3rd. Ed. John Wiley & Sons. **1990**. ISBN 0471510947.

Now, the whole set of descriptors obtained by Saturation Analysis [Fig. 4.9] show that the total number of conformations of CD5 is 2 according to D-I, and 4 according to D-II. Nevertheless, the single conformation [**5s**] –according to D-I & D-II population-ratios- corresponds to three different conformations: [**2Z,S,Z,S**], [**Z,2S,Z,S**], and [**4Z,S**]. This assumption becomes more evident when we just replace on D-II names all the [**Z**] and [**S**] characters with that of [**s**]. This interchange yields again the D-I names and ratios, and support our concern about the inclusion of both distributions within the same class [**s**].

	Conformations from ca05iSA1025					
Descriptor	**I**	**N**	**(%)**	**II**	**N**	**(%)**
Conformations	**5s**	1023	99.80	**2Z,S,Z,S**	607	59.22
				Z,2S,Z,S	407	39.71
				4Z,S	9	0.88
	4s,-	2	0.20	**2Z,S,-,S**	2	0.20
Total	2	1025	100.00	4	1025	100.00

Figure 4.9: CD5 1024 steps Simulated Annealing calculation: Comparison table of both D-I and D-II, and equivalence between full set of descriptors.

Up to this point, it has been proved that D-I does not accomplish the necessary *bijective* correspondence between names –descriptor- and conformations. This *one-to-one* relationship must be kept if we use this index as an indirect help to *"see"* the molecule in space. Then, D-II is finer a tool than D-I to classify conformations but, could there be any unperceived mistake, so far overlooked, as in the case of the D-I definition? PCA results will help in solving this contingency.

Firstly, output data [Fig. 4.10] suggest that only 2 principal components are needed to explain up to 98 % of the system. This means that the original set of 5 variables –flip dihedrals- can be easily reduced to 2 principal components with a loss of information of about 2 %.

	Explained Variance		"Flip" Dihedral				
	indiv.(%)	accumul.(%)	**f1**	**f2**	**f3**	**f4**	**f5**
pc1	**93.334**	93.334	-0.155	0.521	-0.590	0.581	-0.139
pc2	**4.653**	97.987	0.825	-0.206	-0.006	0.460	0.257
pc3	**1.439**	99.426	-0.498	-0.358	0.418	0.651	0.159
pc4	**0.437**	99.863	-0.173	0.144	-0.201	-0.154	0.941
pc5	**0.137**	100.000	0.132	0.733	0.661	0.065	0.063

Figure 4.10: CD5 1024 steps Simulated Annealing calculation: Table of Loadings shows that only 2 principal components are enough to describe almost the whole system.

The plotting of pc1 vs. pc2 [Fig. 4.11] clearly shows 3 apparently different groups of conformations. Nevertheless, further analysis seems to point that, as D-II predicts, the total number of groups is 4 instead: the first one, **A**, is the extremely small set of 2 points down on the left representing conformation [**2Z,S,-,S**]. The second group, **B**, are those 2 enormous areas placed between -20 and -40, consistent with conformation [**2Z,S,Z,S**]. The third group, **C**, is the *"line"* of structures along Y-axis associated to conformation [**4Z,S**], and finally, the fourth group, **D**, is the well-defined area on the right, between 60 and 80, that matches conformation [**Z,2S,Z,S**].

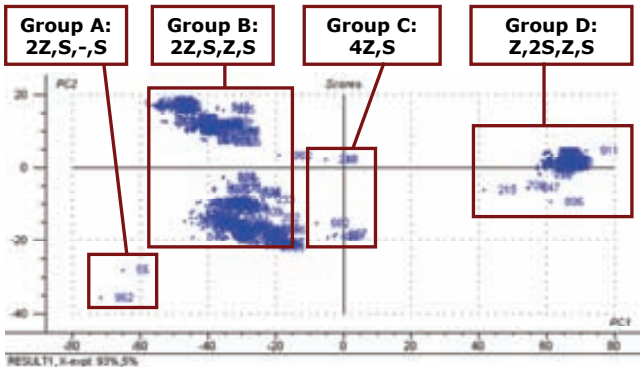


Figure 4.11: CD5 1024 steps Simulated Annealing calculation: Plot of Scores –pc1 vs. pc2- shows 4 conformations.

Anyway, the fact that the name [**2Z,S,Z,S**] was representing two separated groups of extremely well-defined areas was a disturbing result. In principle, this plot seems to point that D-II was not taking into account this potentially important result, meaning that it would have been also incorrectly defined.

Again, under detailed examination of the histograms of the whole set of cyclodextrins [Fig. 4.12], we found 2 possible answers for this inconvenient:

1) The class [-] in both D-I and D-II definitions could be considered to enclose 2 distributions together.
2) The small CD5 has an abnormal behaviour within the area of -15 to 20 degrees: These 2 peaks do not appear when other histograms (6, 7, 8, 14, 21, 26, and 28 glucoses) are examined.

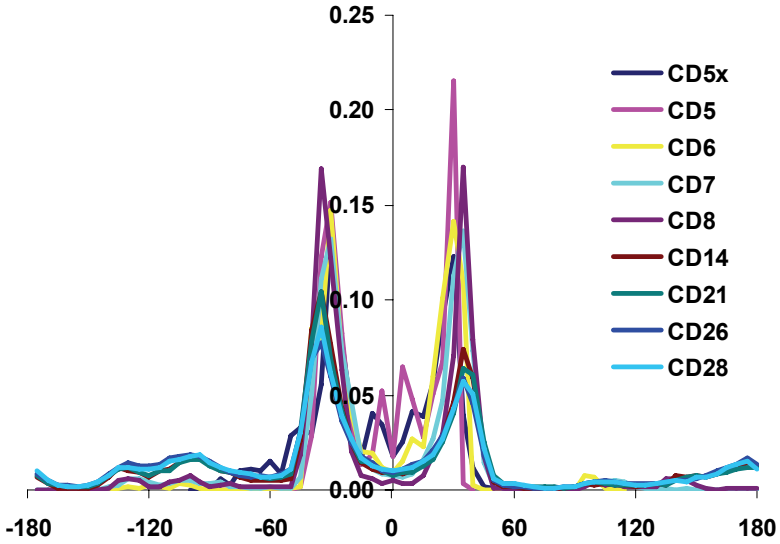


Figure 4.12: Superimposed Histograms of *Flip* Dihedral data obtained from the whole set of Cyclodextrins. Note the special behaviour of ca05x and ca05i in the interval -35 to +35 degrees.

Regarding the first scenario, this possibility was rapidly ruled out because the name [**2Z,S,Z,S**] did not even has any class [-] within the sequence, so the problem cannot be here.

Regarding the second scenario, we thought that the two areas in group II were the consequence of the higher rigidity of the CD5. Ordinary sequences of the type [**Z,Z,Z,Z…**] in CDs other than CD5, involve relative dihedral values of -35 degrees, whilst in those of the type [**S,S,S,S…**], values are about +35 degrees. Therefore, sequences alternating [**Z**]'s with [**S**]'s –as in [**Z,S,Z,S…**]- lead to an average value of 0 degrees. Results obtained support this fact and show that the alignment *all-glucoses-up* is the preferred in small cyclodextrins due to steric effects.

This macrocycle, CD5, contains 5 glucoses, which is an *odd* number of units. Then there are only 2 possible alternate sequences:

1) [**Z,S,Z,S,Z**] that is reduced to [**2Z,S,Z,S**], and
2) [**S,Z,S,Z,S**] that is reduced to [**Z,2S,Z,S**].

In both cases they are forced to accommodate glucose units 1 and 5 trying to fit the "0 degrees average" adjusting the less-stable non-alternate conformations [**…2Z…**] or [**…2S…**] to tighter dihedral values [Fig. 4.13]:

	ca05x	ca05i	caNNi	
[...2Z...]	-10.0	-5.0	-35.0	[...Z...]
[...2S...]	10.0	5.0	35.0	[...S...]

Figure 4.13: Table containing the standard values of [Z] and [S], and their special values in CD5.

In summary, results suggest that the two areas in [**2Z,S,Z,S**] are caused by the 2 distributions mistakenly enclosed respectively under the classes [**Z**] and [**S**] when CD5 is studied [Fig. 4.14]. In fact, the effect in class [**S**] is apparently far smaller than in class [**Z**] according to pc1 vs. pc2 plot, where the group of conformations representing [**Z,2S,Z,S**] is a single group instead of 2 as in the case of [**2Z,S,Z,S**].

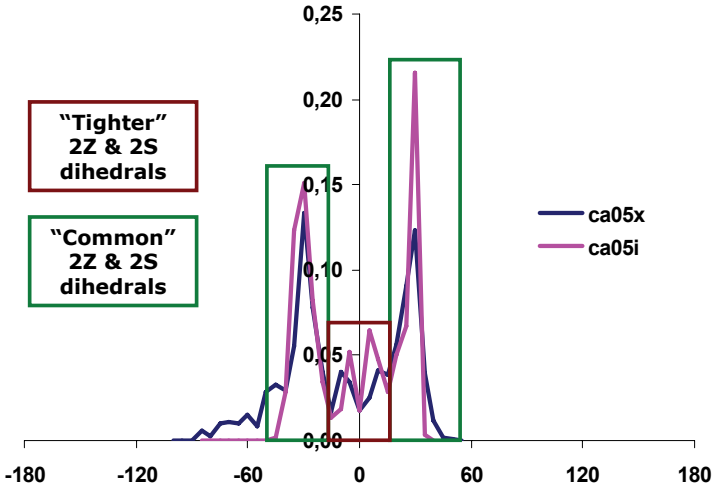


Figure 4.14: Superimposed Histograms of *Flip* Dihedral from ca05x and ca05i. Check the special behaviour in the interval -35 to +35 degrees.

This possibility raises an important question: should we divide [**Z**] and [**S**] into smaller distributions and add new classes to better explain this exceptional behaviour?

This option would lead to the creation of a new and more precise Descriptor, D-III, including 6 or 7 classes, and a better bijective ratio. But the total number of conformations would be significantly increased and the difference between groups of conformations would be reduced. The situation in the case of D-II was critical: we were forced to update the definition because the limits were shifted out of the minima and especially because the problematic class [**s**] enclosed nearly 95% of conformational information in ordinary CDs and about 67% in large CDs [Fig. 4.15].

	Ordinary CD's					Large CD's			
class [s]	**CD5x**	**CD5**	**CD6**	**CD7**	**CD8**	**CD14**	**CD21**	**CD26**	**CD28**
frequency (%)	96.039	99.961	96.699	91.387	90.000	70.411	69.333	63.909	65.561
average	94.817					67.304			
std. dev.	4.076					3.073			

Figure 4.15: Table of probabilities for the whole set of CDs.
Statistical importance of [s] class.

Nevertheless, if we were to use D-III, it would mean that we would be gaining extra difficulties and losing simplicity and easy-handling just to improve our power of prediction exclusively on CD5 a 5% -pc2 explained ratio- from 93% to 98%. Quite a decision.

In this point, a theoretical approach was also considered as an extra source of information.

4.2.5 Estimating Total Number of Conformers: Equations.

Every time the number of classes is extended to allow a better description of the system, the total amount of conformations becomes significantly enlarged. The question is how this increment quantitatively affects the total number of conformations? This section presents the set of equations that correlates the total number of conformations with the number of discrete classes and glucoses.

The idea is, on the one hand, to define a tool that help us to estimate the theoretical total number of conformations not wondering how stable they are. On the other hand, to set a basis that enable a quantitative comparison between D-I, D-II, and D-III in order to take a decision about which of them is the best choice, balancing the explaining ratio to the ease of handling.

First of all, if we were to describe the number of conformations in a linear chain containing 2 glucose units, we realise that only one “flip” dihedral is present [Fig. 4.16]. Suppose that the first glucose unit is fixed in the space and we rotate the second one in order to change conformations. In this situation, although two glucoses are considered, the number of variables is 1 because there is only one flip dihedral angle between them.

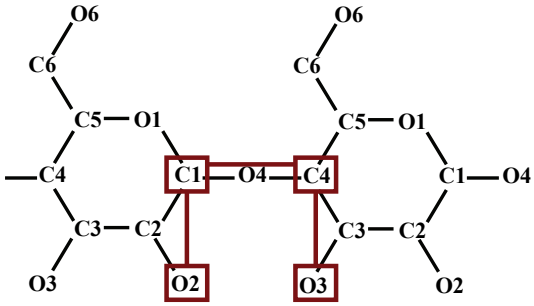


Figure 4.16: 2 glucoses in a linear chain and 1 *flip* dihedral between them.

Now, assume that we are dealing with a linear chain containing 3 glucose units. The procedure is similar to the previous example: The first glucose in the chain is fixed in space and now we rotate the second glucose to allow a conformational change not wondering what happens to the third glucose. When the second glucose has been already placed, the relative rotational position becomes defined by the "flip" dihedral between them, and now we are able to rotate the third glucose keeping previous glucoses fixed in space. This time, 3 glucoses were considered and 2 variables appeared [Fig. 4.17].

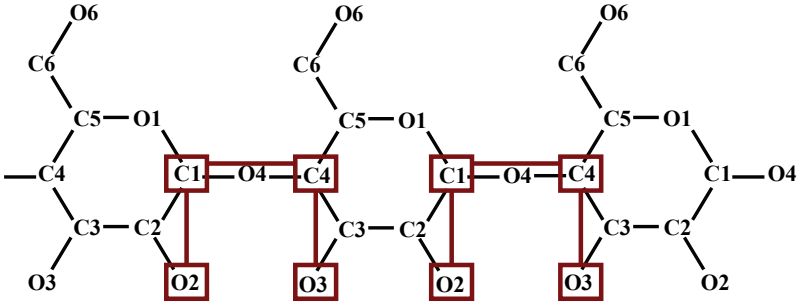


Figure 4.17: 3 glucoses in a linear chain and 2 *flip* dihedrals between them.

The procedure can be repeated the same way in a linear chain containing "n" glucoses and the total number of variables will be "n – 1" because the number of degrees of freedom in this system equals the number of "flip" dihedrals, which also equals the total number of glucoses minus 1.

But cyclodextrins are cyclic macromolecules, the chain of glucoses being closed. Cyclization adds a new "flip" dihedral to the set. *Apparently*, this feature points to the fact that the number of the degrees of freedom is "n", nevertheless, careful examination concludes that this is false.

Say we have CD8, we are changing dihedral values by hand, and then we proceed as in linear chains. We firstly chose any glucose in the CD8 as the glucose number 1. Now, we fix this glucose in space and rotate glucose number 2 to define "flip" dihedral between 1 and 2. Then, once 1 and 2 have been fixed in the space, we rotate glucose number 3 to define "flip" dihedral between 2 and 3. We go on this way until we get to glucose number 8. By the time glucose 8 is to be rotated, all the previous ones in the chain –the other 7- have already been fixed in space. The fact is, when 8 is rotated, since 7 and also 1 are fixed, both "flip" dihedrals –the one between 7 and 8, and the one between 8 and 1- are changed *at the same time*, so the total number of degrees of freedom is not really "n" but "n – 1". Cyclization adds an additional "flip" dihedral but also removes one degree of freedom; so one effect compensates the other.

Now, since every dihedral is staggered into "h" discrete classes, and there are as many free dihedrals as total number of degrees of freedom, "n – 1"; the total number of conformations, in principle, can be expressed in the way:

$$C(h,n) \propto h^{(n-1)}$$

But there is still the question about the cyclic symmetry of cyclodextrins. In order to remove it, we should divide that expression by the total number of glucose units, "n". Finally, the final expression [Eq. 4.1] is found to be:

$$C(k,h,n) = k\frac{h^{(n-1)}}{n}$$

Equation 4.1

Where "k" is the scaling constant, usually set to 1. This function's growing ratio is extremely high [Fig. 418] so a logarithmic form [Eq. 4.2] is also convenient and useful:

$$\ln\left[C(k,h,n)\right] = \ln(k) + (n-1)\ln(h) - \ln(n)$$

Equation 4.2

Descriptor		D-I		D-II		D-III	
Parameter		C(h)	lnC(h)	C(h)	lnC(h)	C(h)	lnC(h)
h		4	4	5	5	6	6
Nº of Glucoses: n	5	51	3.94	125	4.83	259	5.56
	6	171	5.14	521	6.26	1296	7.17
	7	585	6.37	2232	7.71	6665	8.80
	8	2048	7.62	9766	9.19	34992	10.46
	14	4.8E+06	15.38	8.7E+07	18.28	9.3E+08	20.65
	21	5.2E+10	24.68	4.5E+12	29.14	1.7E+14	32.79
	26	4.3E+13	31.40	1.1E+16	36.98	1.1E+18	41.54
	28	6.4E+14	34.10	2.7E+17	40.12	3.7E+19	45.05

Figure 5.18: Table of number of conformations and its logarithm as a function of the number of glucoses and the number of classes.

Now, a few algebraic manipulations render interesting properties. We are interested in the growing ratio of conformations as a function of the number of classes [Eq. 4.3] and [Eq. 4.5] and of glucoses [Eq. 4.4] and [Eq. 4.6]; then, this information is given by the derivative[182] of the last two equations, [Eq. 4.1] and [Eq. 4.2]:

$$\frac{\Delta C(k,h,n)}{\Delta h} \approx \frac{\partial C(k,h,n)}{\partial h} = k\left(\frac{n-1}{n}\right)h^{n-2} = C(k,h,n)\left(\frac{n-1}{h}\right)$$

$$\frac{\Delta C(k,h,n)}{\Delta n} \approx \frac{\partial C(k,h,n)}{\partial n} = C(k,h,n)\left[\left(\frac{n-1}{h}\right) - \frac{1}{n}\right]$$

Equations 4.3 and 4.4

$$\frac{\Delta \ln[C(h,n)]}{\Delta h} \approx \frac{\partial \ln[C(h,n)]}{\partial h} = \left(\frac{n-1}{h}\right)$$

$$\frac{\Delta \ln[C(h,n)]}{\Delta n} \approx \frac{\partial \ln[C(h,n)]}{\partial n} = \ln(h) - \frac{1}{n}$$

Equations 4.5 and 4.6

However, the quotients C(h+1)/C(h) and C(n+1)/C(n) are better parameters to monitor the effect of incrementing the number of classes and glucoses in the total number of conformations:

When the number of classes has been fixed, the amount of conformations rapidly grows [Eq. 4.7] as the number of glucoses is incremented (q = 1, 2, 3…) [Fig. 4.19]. The

[182] Please, note that those two are *approximate* equations and became *nearly* true as the integer numbers "h", and "n" tend to infinite; only in that case the increment Δh behaves as an infinitesimal quantity in comparison to them, and the derivative can be obtained. In fact, derivatives are only properly defined in Real/Complex Functions on the Real/Complex Numbers Field.

impact of the exponential behaviour in conformational search studies is really critical and shows the magnitude of the problem.

$$\frac{C(h,n+q)}{C(h,n)} = \left(\frac{n}{n+q}\right)h^q \Rightarrow q=1 \Rightarrow \frac{C(h,n+1)}{C(h,n)} = \left(\frac{n}{n+1}\right)h$$

Equation 4.7

C(n+1)/C(n)		Number of Classes: h			
		4	5	6	7
N. of Glucoses: n	5	3.33	4.17	5.00	5.83
	6	3.43	4.29	5.14	6.00
	7	3.50	4.38	5.25	6.13
	8	3.56	4.44	5.33	6.22
	14	3.73	4.67	5.60	6.53
	21	3.82	4.77	5.73	6.68
	26	3.85	4.81	5.78	6.74
	28	3.86	4.83	5.79	6.76

Figure 4.19: Table of quotients C(n+1)/C(n) for different number of classes *h*.

Now considering a particular cyclodextrin –fixed number of glucoses, "n"- the effect of incrementing by 1 the number of classes (p = 1) becomes more important when that number is quite low [Eq. 4.8]. I.e. going from 7 to 8 classes is not as bad as going from 4 to 5. Suppose we are dealing with CD28, then in the 7-to-8 case, the ratio C(h+1)/C(h) is nearly 37 while in the 4-to-5 case is almost 414 [Fig. 4.20].

$$\frac{C(h+p,n)}{C(h,n)} = \left(1+\frac{p}{h}\right)^{n-1} \Rightarrow p=1 \Rightarrow \frac{C(h+1,n)}{C(h,n)} = \left(1+\frac{1}{h}\right)^{n-1}$$

Equation 4.8

C(h+1)/C(h)		Number of Classes: h			
		4	5	6	7
N. of Glucoses: n	5	2.44	2.07	1.85	1.71
	6	3.05	2.49	2.16	1.95
	7	3.81	2.99	2.52	2.23
	8	4.77	3.58	2.94	2.55
	14	18.19	10.70	7.42	5.67
	21	86.74	38.34	21.82	14.45
	26	264.70	95.40	47.17	28.17
	28	413.59	137.37	64.20	36.79

Figure 4.20: Table of quotients C(h+1)/C(h) for different number of glucoses *n*.

We need to keep a low number of classes because the number of conformations is quite sensitive to this parameter, therefore we have to carefully evaluate the pros and cons of

incrementing it when this change does not really imply a significant improvement in conformational description.

Therefore, on the basis of the previous results –PCA, histograms and equations- and remembering *Ockham Razor's Principium*[183] our decision was keeping D-II as the chief descriptor in the present work. As said before: *"Good Enough, is Good"*.

4.3 CLASSIFICATION: GEOMETRIC CRITERIA

4.3.1 Data Mining in Conformational Pool

There are two criteria in monitoring a conformational search process; *energetical* and *geometrical*. The first one and probably the most usual in practice according to its simple implementation to any molecule, is seeking for the conformation of minimum energy. A given algorithm continuously generates large amounts of structures and performs a biased filtering favouring selection of the most stable ones. The process is usually terminated when conformations more stable than the best one stored at that time are no longer produced, or when a certain number of steps have been exhausted.

This methodology is suitable for small molecules because their conformational space is often reduced and different conformations usually lead to different energies, and *vice-versa*.

The situation is absolutely different when large macromolecules are considered. Their extremely high flexibility produces large amounts of conformations within a narrow interval of energies [Fig. 4.21]. So the filtering criterion cannot be that of energies because there is no bijective correspondence between energies and conformations.

[183] William of Ockham (**c. 1288 - c. 1348**). *Law of Parsimony: "Entia non sunt multiplicanda praeter necessitatem"*. When multiple competing theories are equal in other respects, the principle recommends selecting the theory that introduces the fewest assumptions and postulates the fewest entities.

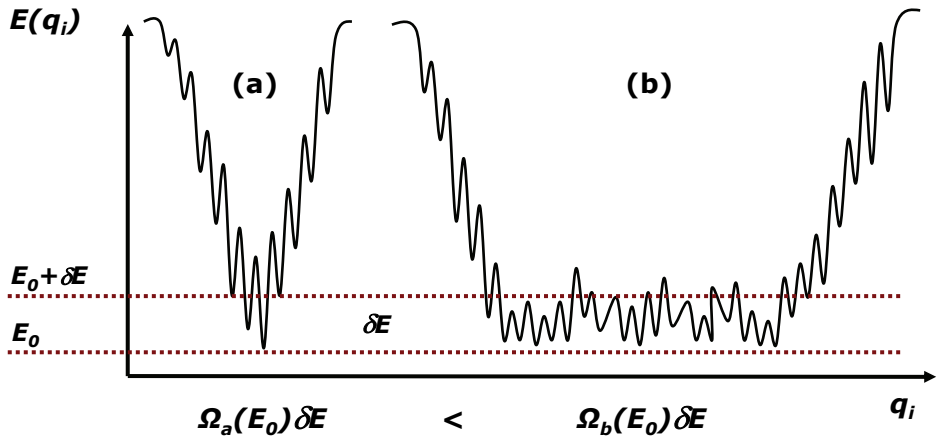


Figure 4.21: Schema of two possible conformational spaces: (a) Typical profile in rigid and small molecules and (b) in large and flexible macromolecules. The number of conformations between E_0 and $E_0 + \delta E$ is often reduced in spaces of the type (a) while in spaces of the type (b) this number is huge.

That is the reason for the second criterion; the geometrical one. Anyway, while the energetical one can be done as the same time as the algorithm produces the set of conformations, the geometrical is applied after the population has been fully obtained. Therefore, you need a conformational pool previous to the analysis. The point is how big this population should be in order to effectively represent the whole system?

Sampling –or selecting the number of steps- is often a problem when you are not sure about the number of conformations to study. Estimating that number for a given macromolecule assumes that you already have certain knowledge about its conformational space, which is precisely what you don't have and what you are trying to obtain. In other words, you are compelled to *"guess"*.

The methodology we are proposing here is one in the structural/geometrical type. It is based on D-I and, especially, D-II nomenclatures, both developed in our group, and analyse/classify conformations automatically taking those descriptors into account as the main Data Mining tool. Besides, one of the goals of this technique is that of supplying information about how well the conformational search has been done, and estimating the necessary number of steps to achieve a worthwhile conformational search.

4.3.2 Conformational Search: Saturation Approach to SA and MD.

4.3.2.1 Deduction

The theoretical basis is quite straightforward, so let us start enunciating the necessary assumptions for deducing this methodology on the basis of the *Markov Chain Mixing Time*:

1) Any given macromolecule –cyclodextrins in our case- is selected.
2) An **n**-step conformational search with Simulated Annealing (SA) algorithm is performed although any other algorithm including a biased roulette wheel proportional selection could be also employed (Maxwell-Boltzmann Factor, proportional selection…).
3) The number of steps, **n** is such an *extremely large* number that fully ensures effective sampling of conformational space: $\{\mathbf{c_n}\}$.

Under such conditions, **k** conformers $\{\mathbf{c_1, c_2, c_3, \ldots, c_k}\}$, $\mathbf{k} \leq \mathbf{n}$, with associated probabilities $\{\mathbf{p_1, p_2, p_3, \ldots, p_k}\}$, being; $\mathbf{p_1} > \mathbf{p_2} > \mathbf{p_3} > \ldots > \mathbf{p_k}$; and $\Sigma\mathbf{p_k} = 1.0$, are found [Fig. 4.22].

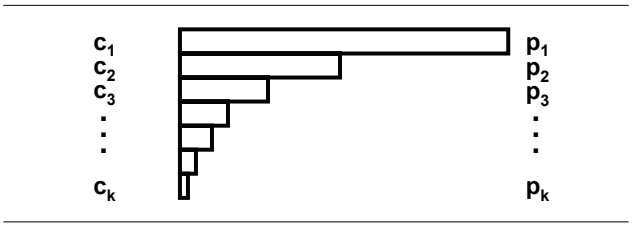


Figure 4.22: Schema of conformations and probabilities sorted by decreasing order.

Now, say we perform another SA conformational search for the same molecule under the same conditions changing *only* the total number of steps from **n** to **2n**, so a new conformational pool is generated: $\{\mathbf{c_{2n}}\}$.

If the **n**-step calculation was effectively done the first time and the conformational space was so exhaustively explored, no other different conformations will be found in the **2n**-step set $\{\mathbf{c_{2n}}\}$; therefore the total number of conformations, **k**, will remain unaltered.

Furthermore, since the number of steps in the first search was extremely large, the energy-biased conformational selection was statistically representative and the probabilities for every conformer were then correctly established. This means that the **2n**-step calculation will not promote any improvement/change in the probability ratios and therefore the set of values, $\{\mathbf{p_1, p_2, p_3, \ldots, p_k}\}$, will also remain unaltered.

4.3.2.2 Conclusion

For a given macromolecule, there exists a critical number of steps called *Mixing Time*, **N**, that is a lower limit for effective conformational search. Therefore, any calculation involving **M** steps, where $\mathbf{N} \leq \mathbf{M} < \infty$ will led to **k** conformers with probabilities $\{\mathbf{p_1, p_2, p_3, \ldots, p_k}\}$. And then, the quantities:

1) Total number of conformers, **k,** and
2) Their probabilities $\{\mathbf{p_1, p_2, p_3, \ldots, p_k}\}$

Can be considered as *invariants* in any conformational search process involving **M** steps. This is the starting point.

4.3.2.3 Data Mining: Saturation Diagram

In the end, the problem in the geometrical approach is estimating **N** because we either cannot predict its value, or it is an unaffordable quantity. We finally realised that although **N** can be considered unreachable in most cases, we can *approach* its value *step by step*; that is really important because the asymptotic behaviour supplies valuable information:

1) We can estimate how well the conformational search evolved and,
2) How close to the **N** real value we are.

In this way, the idea is performing *series of SA calculations* where the total number of steps is *systematically incremented* in every new calculation:

1) *Arithmetically*. [Eq. 4.9]. I.e. series of SA calculations involving 1000; 2000; 3000, 4000… steps (a = 1000):

$$C(i) = ai$$

Equation 4.9

2) *Geometrically*. [Eq. 4.10]. I.e. series of SA calculations where the number of steps is 1; 2; 4; 8; 16; 32; 64… steps (a = 1; z = 2):

$$C(i) = az^{i}$$

Equation 4.10

In the present work the geometrical series were selected because the information ratio [Eq. 4.11] is always maintained constant at 50% [Eq. 4.12], while in the arithmetic series is hyperbolically reduced [4.13] as the number of iterations grows [Fig. 4.23]:

$$I(i) = \left[\frac{C(i) - C(i-1)}{C(i)}\right]$$

$$I^{geom}(i) = \left[\frac{az^{i} - az^{i-1}}{az^{i}}\right] = \left(1 - z^{-1}\right)$$

$$I^{arith}(i) = \left[\frac{ai - a(i-1)}{ai}\right] = \frac{1}{i}$$

Equations 4.11; 4.12 and 4.13

As the iterations grow, the number of conformations necessary for the saturation analysis becomes unaffordable [Eq. 4.14] and [Eq. 4.15], especially in the case of the geometrical approach [Fig. 4.23]:

$$S^{geom}(i) = \sum_{i=0}^{k} az^{i} = \frac{a\left(1 - z^{k+1}\right)}{1 - z}$$

$$S^{arith}(i) = \sum_{i=1}^{k} ai = \frac{ka(1 + k)}{2}$$

Equations 4.14 and 4.15.

Iteration		0	1	2	3	4	5	6	7	8	9	10
Arith.	Steps	1	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
	Info.	100.0	99.9	50.0	33.3	25.0	20.0	16.7	14.3	12.5	11.1	10.0
	Sum	0	1000	3000	6000	10000	15000	21000	28000	36000	45000	55000
Geom.	Steps	1	2	4	8	16	32	64	128	256	512	1024
	Info.	100.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
	Sum	1	3	7	15	31	63	127	255	511	1023	2047

Figure 4.23: Table of comparison between arithmetical and geometrical approaches to conformational search.

This is a serious drawback because you are forced to study different series of data for a single macromolecule involving several files and large amounts of descriptors. In the end, it is uncomfortable, computationally time consuming and, in principle, people would not adopt such a methodology. Fortunately, a solution was found.

The basic assumption is that the SA algorithm creates an ensemble in thermodynamical equilibrium once you have reached a large amount of steps. This point can be discussed as the SA process reproduces a *Markov Chain* of states [Fig. 4.24] and therefore *memory* is only partially retained in consecutive conformational transitions between neighbours depending on the conditions of the SA algorithm.

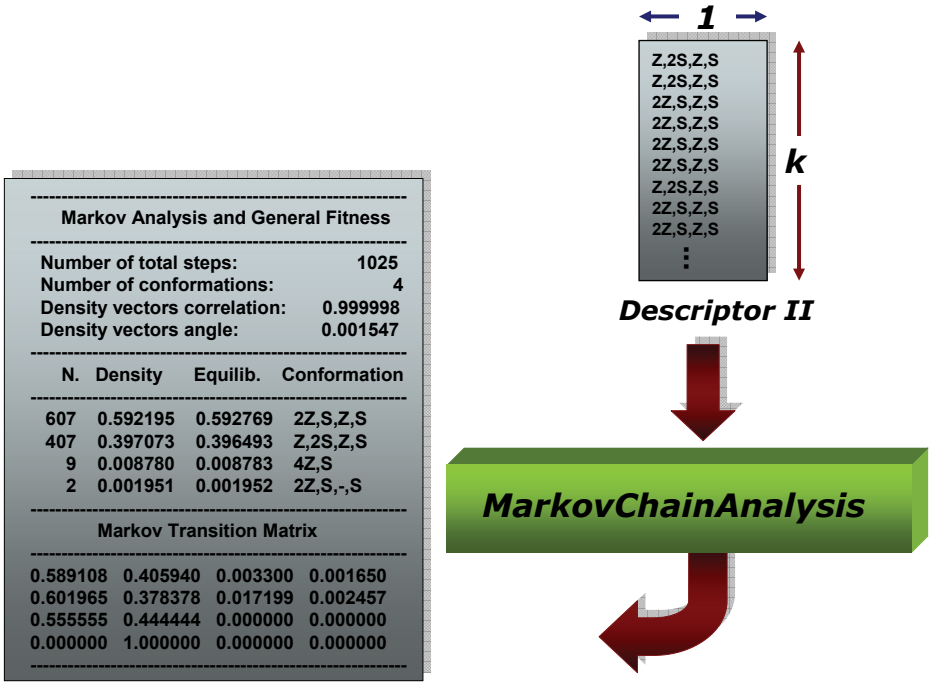


Figure 4.24: Flow chart of *MarkovChainAnalysis.exe*. The input files –any of them both- contains the *k* x *1* array of descriptors created by *NameGiantCyD*, where *k* is the number of conformations. The program processes the file and produces an output file containing the statistical weights and the Markov Transition Matrix.

Anyway, the underlying idea is, assuming we have a statistically representative conformational ensemble, *not to perform all the calculations in the series, but only the last one*. Instead of doing "i" calculations; the 1-step SA, the 2-step SA, the 4-step SA… and the 2^i-step SA, just do the 2^i-step SA and analyse it by parts. The single file containing the vertical array of, say, 1024 molecular descriptors –conformations-, stored in the same order in which they were produced, is iteratively processed beginning from the head and selecting each time as many lines –descriptors- as the total number of conformations that would have resulted from a calculation at iteration "i" in the series if it had been done. The process goes on, then results get improved and become statistically better as the number of conformations studied grows and the whole file is analysed [Fig. 4.25].

ca05iSA1025 ratios (N)													
Iteration [i]		0	1	2	3	4	5	6	7	8	9	10	**10***
Steps [2^i]		1	2	4	8	16	32	64	128	256	512	1024	**1025**
Conformations [k]		1	1	2	2	2	2	3	4	4	4	4	**4**
1	**2Z,S,Z,S**			2	5	11	21	37	71	152	301	606	**607**
2	**Z,2S,Z,S**	1	2	2	3	5	11	26	55	101	203	407	**407**
3	**4Z,S**								1	2	7	9	**9**
4	**2Z,S,-,S**							1	1	1	1	2	**2**

ca05iSA1025 ratios (%)													
Iteration [i]		0	1	2	3	4	5	6	7	8	9	10	**10***
Steps [2^i]		1	2	4	8	16	32	64	128	256	512	1024	**1025**
Conformations [k]		1	1	2	2	2	2	3	4	4	4	4	**4**
1	**2Z,S,Z,S**	0.0	0.0	50.0	62.5	68.8	65.6	57.8	55.5	59.4	58.8	59.2	**59.2**
2	**Z,2S,Z,S**	100.0	100.0	50.0	37.5	31.3	34.4	40.6	43.0	39.5	39.6	39.7	**39.7**
3	**4Z,S**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.8	1.4	0.9	**0.9**
4	**2Z,S,-,S**	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.8	0.4	0.2	0.2	**0.2**

Figure 4.25: Tables of saturation analysis –in number of conformations and percentage- for CD5 1024 steps Simulated Annealing calculation.

The plot of the probability ratios against the number of steps is the final source of information in our conformational search [Fig. 4.26].

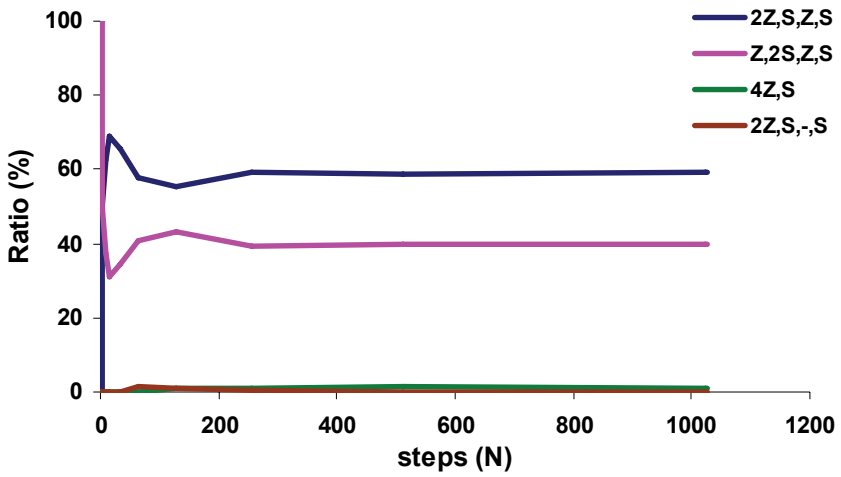


Figure 4.26: Graphic of saturation analysis for CD5 1024 steps Simulated Annealing calculation. In this plot –ratio vs. number of steps-, conformations represented by coloured lines evolve until their ratios become constant –slope equals zero-: that is the *saturation point*.

This process is automatically done with the software ***statistics.exe***. Briefly, this script receives an input file containing the vertical array of descriptors and produces two output files containing the quadratic analysis and the summary [Fig. 4.27]. The full code is included, as in the previous cases, in the Appendix section; Chapter 11.

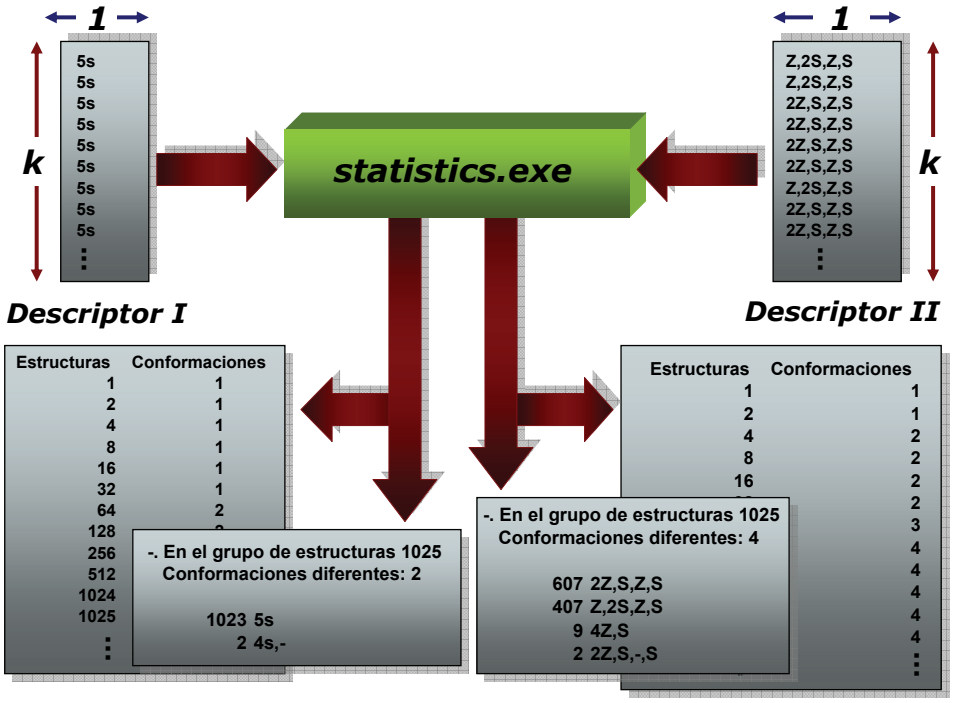


Figure 4.27: Flow chart of *statistics.exe*. The input files –any of them both- contains the *k* x *1* array of descriptors created by *NameGiantCyD*, where *k* is the number of conformations. The program processes the file and produces 2 output files: the SHORT file, containing the summary, and the LONG file, containing every step in the quadratic saturation analysis.

4.3.3 Study of Stability: Trajectories Overlapping Ratio in MD.

When a Molecular Dynamics –MD- calculation is carried out, as in the case of Conformational Search, you cannot be sure about how efficiently the exploration of conformational space has been done. The problems like those of estimating the right simulation times, the sampling ratios and the overall good-of-fitness of the simulation are still present.

One the drawbacks related to large macromolecules, especially in solvent calculations, is evaluating how far from the starting position is able to evolve the molecule all over the surrounding areas within conformational space. It is supposed that a good MD simulation should allow a full sampling, this process being partially biased towards the best minima. Only when this condition is accomplished it is possible to ensure full prediction.

So, what happens when you don't really know how efficiently the MD was done? How about the results? The sampling? The simulation times? The validity of the predictions?

This section proposes a new approach to MD and also offers a monitoring technique for evaluating how well the calculation was done.

The underlying idea is related to that one of Monte Carlo - Stochastic Dynamics (MC/SD) methodology as explained in MacroModel 9.0 User Manual[184]:

MC/SD methodology[185] alternates between MC and SD methods to produce a trajectory. The ratio SD-steps/MC-steps can be adjusted to best fit our objectives: The point is that after a given number of SD steps an MC iteration is performed, so the current conformation is moved farther in conformational space. The important improvement in calculation times relies on the fact that sampling the energetic landscape this way is rapidly done.

[184] **MacroModel 9.0**. *User Manual. Revision A, March 2005*. Ed. Schrödinger Press. **2005**.
[185] (a) Guarnieri, F.; Clark Still, W.; *J. Comput. Chem*. **1994**. *15(11)*. 1302-1310. (b) Senderowitz, H.; Guarnieri, F.; Clark Still, W.; *J. Am. Chem. Soc*. **1995**. *117(31)*. 8211-8219.

As previously said, Monte Carlo –and Genetic Algorithms- were not fully suitable for our research, nevertheless, on the basis of our experience we found that better answers could be obtained running several "short" MD calculations starting from different conformations -obtained by means of conformational search techniques, as in the present thesis-, rather than running a single, longer MD calculation starting from a single conformation, which is in principle a procedure quite similar to the one used in the MC/SD Method.

4.3.3.1 Deduction

Let us start enunciating the necessary assumptions in the theoretical basis:

1) A total number of MD calculations, **r**, will be run.
2) Simulation times and sampling ratios will be held constant; therefore, **n** conformations under equal conditions will be obtained in each MD.
3) Starting conformations can be selected by their statistical weights or just randomly. This point becomes irrelevant when sampling is representative.
4) The **r** trajectories are analysed on the basis of geometrical criteria; D-I, D-II, and the flip dihedral.

Then, once we have the results from the set of trajectories [Fig. 4.28], we compare them. In principle, just to explain the procedure, we will focus on the whole file containing the full array of descriptors from every MD rather than in the quadratic-iterative analysis, although this information is also of use.

Now, let us consider the 3 statistical possibilities. Just to allow better comprehension, examples will be discussed assuming a total number of 3 MD calculations (**r** = 3).

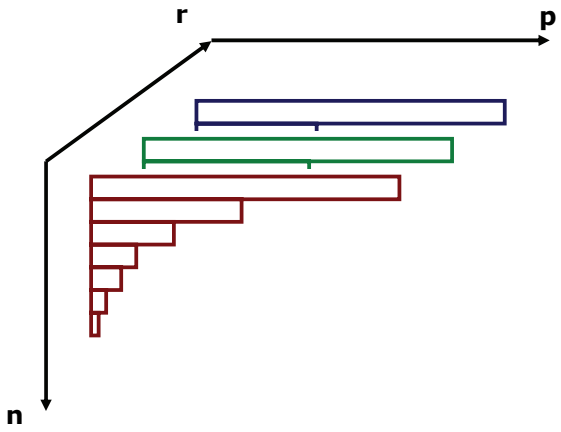


Figure 4.28: Matrix *n* x *r* x *p* containing the data from a series of MD calculations, being *n* the number of conformations, *p* their probabilities and *r* the number of trajectories/MD considered in this analysis.

4.3.3.1.1 Full Conformational Space Overlapping –near 100%-.

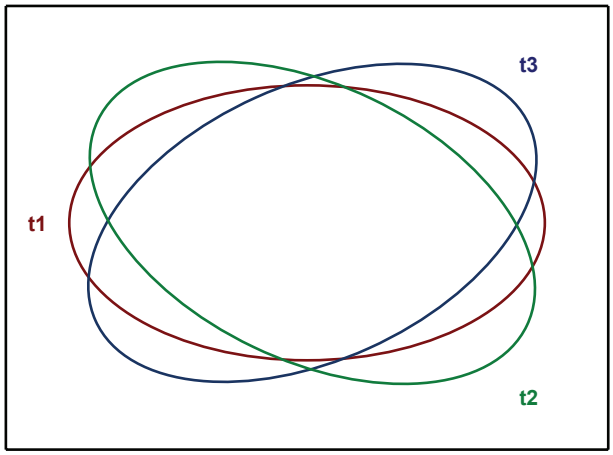


Figure 4.29: Schema of a conformational space being sampled by 3 MD trajectories. The 3 of them overlap almost completely and cover nearly the full conformational space: good sampling.

This situation [Fig. 4.29] is often infrequent. Only small molecules fall usually in this category. The properties in this set of MD include:

1) Effective Exploration: The whole conformational space is covered.
2) Full Sampling: The 3 MD calculations move over the same conformational space area.
3) The total number of conformers found in each trajectory is the same; **k** = **l** = **m**, and equals the total number of conformations, **k**.

$$MD(r=1)=\begin{bmatrix} {}^{1}c_1 \\ {}^{1}c_2 \\ {}^{1}c_3 \\ \vdots \\ {}^{1}c_k \end{bmatrix}; MD(r=2)=\begin{bmatrix} {}^{2}c_1 \\ {}^{2}c_2 \\ {}^{2}c_3 \\ \vdots \\ {}^{2}c_l \end{bmatrix}; MD(r=3)=\begin{bmatrix} {}^{3}c_1 \\ {}^{3}c_2 \\ {}^{3}c_3 \\ \vdots \\ {}^{3}c_m \end{bmatrix}$$

4) The happening ratios of every conformation in each MD trajectory –sorted in decreasing order- are kept constant and are the same in every MD equivalent position. I.e.

$$\begin{array}{ccccc} {}^{1}p_1 & = & {}^{2}p_1 & = & {}^{3}p_1 \\ {}^{1}p_2 & = & {}^{2}p_2 & = & {}^{3}p_2 \\ {}^{1}p_3 & = & {}^{2}p_3 & = & {}^{3}p_3 \\ \vdots & & \vdots & & \vdots \\ {}^{1}p_k & = & {}^{2}p_k & = & {}^{3}p_k \end{array}$$

5) All four previous properties lead to the extremely important corollary: Once the set of conformations in every MD have been sorted, equivalent conformations occupy equivalent positions in the array:

$$\begin{array}{ccccc} {}^{1}c_1 & = & {}^{2}c_1 & = & {}^{3}c_1 \\ {}^{1}c_2 & = & {}^{2}c_2 & = & {}^{3}c_2 \\ {}^{1}c_3 & = & {}^{2}c_3 & = & {}^{3}c_3 \\ \vdots & & \vdots & & \vdots \\ {}^{1}c_k & = & {}^{2}c_k & = & {}^{3}c_k \end{array}$$

Property number 5 is the theoretical basis of our overlapping analysis, and it will be shown in detail later. Now, let us consider a totally different statistical situation.

4.3.3.1.2 No Conformational Space Overlapping –0%-.

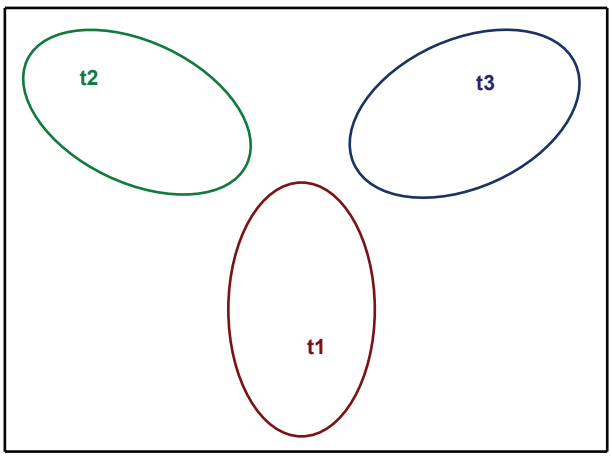


Figure 4.30: Schema of a conformational space being sampled by 3 MD trajectories. There is no overlapping between trajectories and the conformational space is poorly explored: bad sampling.

This is the common scenario in large macromolecules [Fig. 4.30] and, unfortunately, is quite a frequent one. The properties in this set of MD include:

1) Ineffective Exploration: Only small parts of the conformational space are covered. In the presence of other areas not explored, it would be impossible to determine their importance ratio. In summary; full uncertainty.
2) Incomplete Sampling: The 3 MD calculations move over different areas in conformational space. There is no trajectory overlapping so no common areas are sampled.
3) Since the 3 MD evolve over different areas of the conformational space, the conformations found in every MD are all different to one another. Therefore the number of conformers found in each trajectory is expected to be different; $\mathbf{k} \neq \mathbf{l} \neq \mathbf{m}$, and accomplish that the total number of conformations found jointly in the 3 MD is $\mathbf{g} = \mathbf{k} + \mathbf{l} + \mathbf{m}$.

$$MD(r=1)=\begin{bmatrix} {}^{1}c_1 \\ {}^{1}c_2 \\ {}^{1}c_3 \\ \vdots \\ {}^{1}c_k \end{bmatrix}; MD(r=2)=\begin{bmatrix} {}^{2}c_1 \\ {}^{2}c_2 \\ {}^{2}c_3 \\ \vdots \\ {}^{2}c_l \end{bmatrix}; MD(r=3)=\begin{bmatrix} {}^{3}c_1 \\ {}^{3}c_2 \\ {}^{3}c_3 \\ \vdots \\ {}^{3}c_m \end{bmatrix}$$

$$
\begin{array}{ccccc}
{}^{1}c_1 & \neq & {}^{2}c_1 & \neq & {}^{3}c_1 \\
{}^{1}c_2 & \neq & {}^{2}c_2 & \neq & {}^{3}c_2 \\
{}^{1}c_3 & \neq & {}^{2}c_3 & \neq & {}^{3}c_3 \\
\vdots & & \vdots & & \vdots \\
{}^{1}c_k & \neq & {}^{2}c_l & \neq & {}^{3}c_m
\end{array}
$$

4) Obviously, the happening ratios of the conformations in the MD trajectories are expected to be different. I.e.

$$
\begin{array}{ccccc}
{}^{1}p_1 & \neq & {}^{2}p_1 & \neq & {}^{3}p_1 \\
{}^{1}p_2 & \neq & {}^{2}p_2 & \neq & {}^{3}p_2 \\
{}^{1}p_3 & \neq & {}^{2}p_3 & \neq & {}^{3}p_3 \\
\vdots & & \vdots & & \vdots \\
{}^{1}p_k & \neq & {}^{2}p_k & \neq & {}^{3}p_k
\end{array}
$$

And now, the intermediate situation between 100% and 0% overlapping ratios will be discussed.

4.3.3.1.3 Partial Conformational Space Overlapping –Q%-.

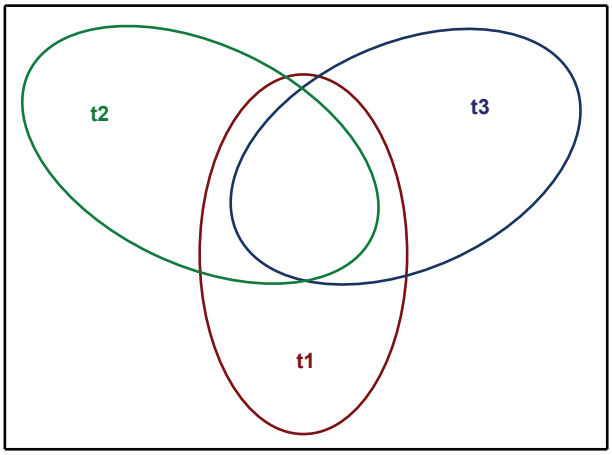


Figure 4.31: Schema of a conformational space being sampled by 3 MD trajectories. There is a partial side overlapping between trajectories and the conformational space is incompletely explored: partial sampling.

This is the common situation in medium-sized macromolecules [Fig. 4.31] and is also quite frequent. The properties in this set of MD include:

1) Partially Effective Exploration: The whole conformational space could be regarded as satisfactorily well explored, but only when the 3 MD are jointly considered.
2) Partial Sampling: The conformational space is covered with partial overlapping between the 3 MD trajectories. It is assumed that the common area is the place where the most stable conformations are located.
3) When comparing conformations between the 3 MD, it is found that some of them are different, and some of them are common to the group. So, the number of conformers found in each trajectory is expected to be different; $\mathbf{k} \neq \mathbf{l} \neq \mathbf{m}$, and the total number of conformations in the 3 MD is $\mathbf{x} < \mathbf{k} + \mathbf{l} + \mathbf{m}$.

$$MD(r=1) = \begin{bmatrix} {}^1c_1 \\ {}^1c_2 \\ {}^1c_3 \\ \vdots \\ {}^1c_k \end{bmatrix}; MD(r=2) = \begin{bmatrix} {}^2c_1 \\ {}^2c_2 \\ {}^2c_3 \\ \vdots \\ {}^2c_l \end{bmatrix}; MD(r=3) = \begin{bmatrix} {}^3c_1 \\ {}^3c_2 \\ {}^3c_3 \\ \vdots \\ {}^3c_m \end{bmatrix}$$

$$\begin{array}{ccccc} {}^1c_1 & = & {}^2c_1 & = & {}^3c_1 \\ {}^1c_2 & = & {}^2c_2 & = & {}^3c_2 \\ {}^1c_3 & \neq & {}^2c_3 & \neq & {}^3c_3 \\ \vdots & & \vdots & & \vdots \\ {}^1c_k & \neq & {}^2c_l & \neq & {}^3c_m \end{array}$$

4) The happening ratios of the conformations in each MD trajectory are expected to be different irrespective of the fact that they are different or common conformations. I.e.

$$\begin{array}{ccccc} {}^1p_1 & \neq & {}^2p_1 & \neq & {}^3p_1 \\ {}^1p_2 & \neq & {}^2p_2 & \neq & {}^3p_2 \\ {}^1p_3 & \neq & {}^2p_3 & \neq & {}^3p_3 \\ \vdots & & \vdots & & \vdots \\ {}^1p_k & \neq & {}^2p_k & \neq & {}^3p_k \end{array}$$

Now that the three possibilities have been shown, the overlapping analysis will be discussed, in order to get a quantitative answer to conformational exploration.

4.3.3.2 Discussion and Equations

The three previous cases have been shown as if they were *uncorrelated* scenarios for different-sized molecules when, in fact, the only truth is that they are just the *same* and the only difference is just as *a matter of time* [Fig. 4.32].

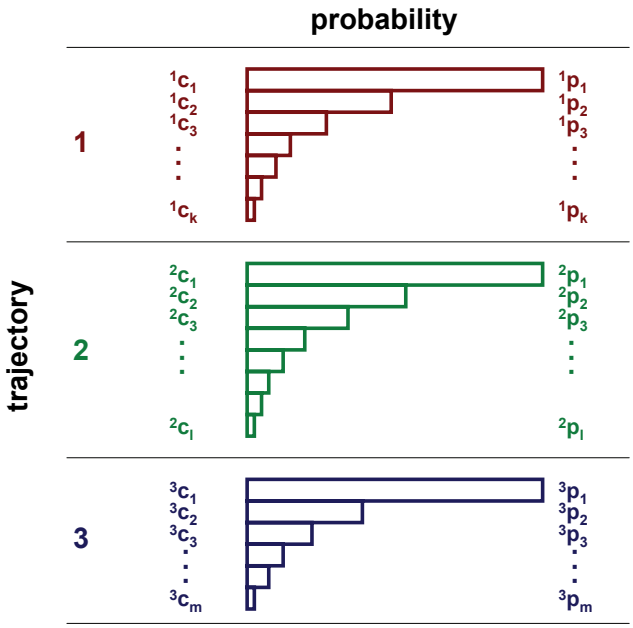


Figure 4.32: Conformational analysis of 3 MD trajectories. Both conformations *c* and probabilities *p* are doubled indexed: the *top-left* one refers to the trajectory and the *bottom-right* one refers to the conformation already sorted by probability.

Irrespective of the molecular size, when a series of MD calculations are performed as in the present case, the evolution of the system is clear: At the beginning, the areas covered by the MD calculations are unconnected and therefore the behaviour is similar to that case of *no overlapping*. As time goes by, the individual areas explored by every MD calculation spread all over conformational space; then trajectories begin to overlap as in the case of *partial overlapping*. In the end, if the system were given infinite time, the individual MD trajectories would cover up the full conformational space, as in the case of *full overlapping*. Therefore, the overlapping ratio might be taken as an index for evaluating how effectively the exploration of the conformational space has been done.

4.3.3.2.1 Overlapping Equation I: Index ω

Then, let us outline the deduction of *Equation I*. For better comprehension, we will assume that 3 MD calculations under similar conditions were done and only the best conformation will be considered.

We firstly calculate the ratios of the best conformation, $\mathbf{c_1}$, in every MD. Then we sum up the three files containing the array of descriptors into a new one [Fig. 4.33] and then we calculate its ratio in the whole set of MD calculations:

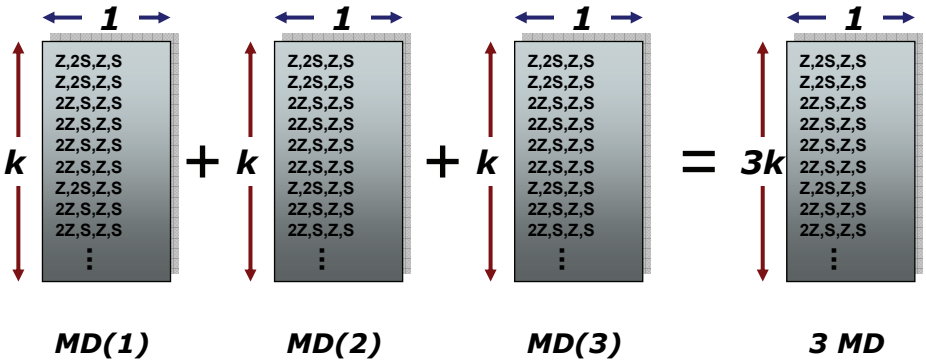


Figure 4.33: The single files containing the array of *k* x *1* descriptors of MD trajectories are concatenated together using the UNIX/Linux *cat* command. The result is a new file containing all the descriptors that will also be processed by *statistics.exe* C-shell script.

Say we are in the case of 100% overlapping. In this situation, the ratio in MD(r = 1) is the same to that one in MD(r = 2) and also equals the one in MD(r = 3). But, due to the fact that we are normalizing by the total number of conformations, the ratio in the file containing the 3 MD descriptors will be also the same.

$$^{1}p_1 = {}^{2}p_1 = {}^{3}p_1 = {}^{full}p_1$$

Now, suppose we are in the case of 0% overlapping. In this situation, since trajectories do not share any conformation in common, the best conformation in, say **t1**, will not appear in any of the other two MD. Then the ratio in the file containing the 3 MD descriptors will be that one in **t1** divided by 3.

$$^{1}p_1 \neq {}^{2}p_1 = {}^{3}p_1 = 0 \Rightarrow {}^{full}p_1 = \frac{^{1}p_1}{r}$$

Then, the *maximum distance ratio*, **d**, for conformation **j** in MD calculation **i** in a group of **r** MD calculations can be estimated by [Eq. 4.16]:

$$^{i}d_j = {}^{i}p_j - \frac{^{i}p_j}{r} = {}^{i}p_j\left(1 - \frac{1}{r}\right)$$

Equation 4.16

Let us consider the case of partial overlapping. We now define the *individual distance ratio* of conformation **j** in MD **i** by the expression [Eq. 4.17]:

$$ {}^{i}v_j = \left| \frac{\sum_{i}^{r} {}^{i}p_j}{r} - \frac{{}^{i}p_j}{r} \right| = \left| \left\langle p_j \right\rangle - \frac{{}^{i}p_j}{r} \right| $$

Equation 4.17

Now let us check the value of ${}^{i}v_j$ in the cases of 100% and 0% overlapping. In the first case, according to ${}^{1}p_1 = {}^{2}p_1 = {}^{3}p_1 = {}^{full}p_1$ it is easy to realise that the average $\left\langle p_j \right\rangle$ is exactly the same to ${}^{1}p_1 = {}^{2}p_1 = {}^{3}p_1$, and then ${}^{i}v_j = {}^{i}p_j\left(1 - 1/r\right)$. In the second case, where the conformation under study appears only in one trajectory ${}^{1}p_1 \neq {}^{2}p_1 = {}^{3}p_1 = 0$ the average probability comes to be $\left\langle p_j \right\rangle = {}^{1}p_1/3$ and then ${}^{1}v_j = 0$. In summary, this information helps to introduce the *Overlapping Equation I* [Eq. 4.18] as defined:

$$ {}^{i}\omega_j\left[{}^{i}p_j\right] = \left(\frac{{}^{i}v_j}{{}^{i}d_j} \right) $$

Equation 4.18

Where it is easy to check that, ${}^{i}\omega_j$ = 1.0 in the full overlapping case, ${}^{i}\omega_j$ = 0.0 in the case of no overlapping, and 0.0 < ${}^{i}\omega_j$ < 1.0 in the partial overlapping case.

The set of indexes ${}^{i}\omega_j$ are numerical quantities that represent the overlapping ratios between trajectories and, therefore, can be regarded as a useful tool to monitor the effectiveness of a conformational space exploration in MD.

4.3.3.2.2 Overlapping Equation II: Index λ

A similar equation based on the number of conformations in every MD calculation can also be obtained under similar deductions. It was shown that, in the case of full overlapping, the total number of conformers in every MD was **k** = **l** = **m**. While in the

case of no overlapping, this number was **g** = **k** + **l** + **m**. Then, the partial overlapping case was shown to be an intermediate situation where the total number of conformers was **x** < **k** + **l** + **m**.

Then, suppose that the number of conformers found in MD **i** is represented by ${}^{i}n$, the average number of conformers is $\langle n \rangle$, and the total maximum number of conformers in the set of MD calculations, **q**, is then defined by [Eq. 4.19]:

$$q = \sum_{i}^{r} {}^{i}n$$

Equation 4.19

Now, as in Equation I, we sum up the three files into a single file containing the array of descriptors and we calculate the *real* total number of different conformations, **t**, with the script ***statistics.exe***. Combining all of these parameters we are able to define the *Overlapping Equation II* [Eq. 4.20]:

$$\lambda\left[{}^{i}n\right] = \left(1 - \frac{t - \langle n \rangle}{q - \langle n \rangle}\right) = \left(\frac{q - t}{q - \langle n \rangle}\right)$$

Equation 4.20

In which it is easy to check that, $\lambda = 1.0$ in the full overlapping case, $\lambda = 0.0$ in the case of no overlapping, and $0.0 < \lambda < 1.0$ in the partial overlapping case.

4.4 INTEGRATED PROTOCOL FOR CONFORMATIONAL SEARCH.

In summary, our proposal for studying large macromolecules requires a 2-stage process, involving a Conformational Search followed by a Molecular Dynamics study. Both methodologies are well-known ones in the Molecular Modelling area of knowledge; nevertheless, the main contribution made by this research work is the development of suitable methodological techniques –saturation analysis and trajectory overlapping ratio- for monitoring quantitatively how SA and MD calculations evolve.

Step I. Conformational Search

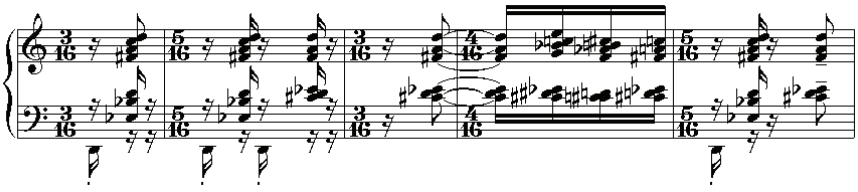
1) Define the appropriate geometrical molecular descriptor
    - Mathematical study (Histograms, Distributions, PCA…) of molecular parameters (geometric, energetic, topologic…).
2) Generate an ensemble by means of Simulated Annealing
3) Check convergence:
    - Descriptor: ***Saturation Analysis***
    - Markov Matrix
4) Analysis: obtain molecular information from the ensemble in equilibrium

Step II. Molecular Dynamics

1) Extract some representative conformations from Conformational Search
2) Run MD calculations under the same conditions starting from those conformations
3) Check MD Convergence:
    - Descriptor **index λ**: ***Trajectory Overlapping Ratio***
    - Saturation of the Total Number of Conformations vs. time
4) Analysis: obtain molecular dynamical information from the trajectories

NOTE: The saturation tools developed in the present thesis could be also of use when applied in the analysis of different sets of ensembles or trajectories obtained by other methodologies like *parallel tempering*[186], *replica exchange method*[187] (REM), or *umbrella sampling*[188]. This assumption has not been yet proved; however, it seems to be a good path for future projects.

[186] D.D. Frantz, D.D.; Freeman D.L.; Doll, J.D.; *J. Chem. Phys*. **1990**. *93(4)*. 2769-2784.
[187] Hukushima K.; Nemoto, K.; *J. Phys. Soc. Jpn*. **1996**. *65*. 1604-1608.
[188] Valleau, J.P.; *J. Chem. Phys*. **1993**. *99(6)*. 4718-4728.



Igor Stravinsky
The Rite of Spring
(Pictures from Pagan Russia)
Ballet in two parts
(1913)
Sacrificial Dance (The Chosen One)

"Searching for patterns is such a hard work..."

5 RESULTS II: CONFORMATIONAL SEARCH

5.1 CONFORMATIONAL SEARCH:

This chapter contains the first part of the research, the Conformational Search. Within this frame, two objectives were established:

- Test the efficiency of SA methodology in combination with Saturation Analysis
- Explore the Conformational space of the CDs selected as benchmark and do a comparative study of their conformational properties.

The last objective will allow selecting the most representative conformers to carry out the subsequent MD studies necessary to prove/false MD dependence on starting conformation.

5.2 SIMULATED ANNEALING (THERMAL SHOCK APPROACH)

Simulated Annealing (see appendix: chapter IX methodology) has been applied over the entire group of CDs:

- Macrorings including the set of RESP-charges calculated by Dr. Iván Beà and also used by Dr. Itziar Maestre and Dr. Miguel de Federico:
  - ❖ Small cyclodextrins: CD5, CD6, CD7 and CD8.
  - ❖ Large cyclodextrins: CD14, CD21, CD26 and CD28.

- The macroring including the set of RESP-charges calculated by Dr. Javier Pérez:
  - ❖ Small cyclodextrin CD5x (where "x" stands for the new set of charges).

5.2.1 Slow SA [300 ps]. 1024 steps

The first series of 1024-steps SA calculations were carried out following the schema in [Fig. 5.1]. Every step included *"in vacuo"* conditions, 300 picoseconds, and external constraints to avoid chair-to-boat conformational interchanges in glucoses. Detailed

information regarding *slow* SA conditions and further computational aspects can be found in the DVD: appendix 10.1.1.1 (chapter X, AMBER 7 files).

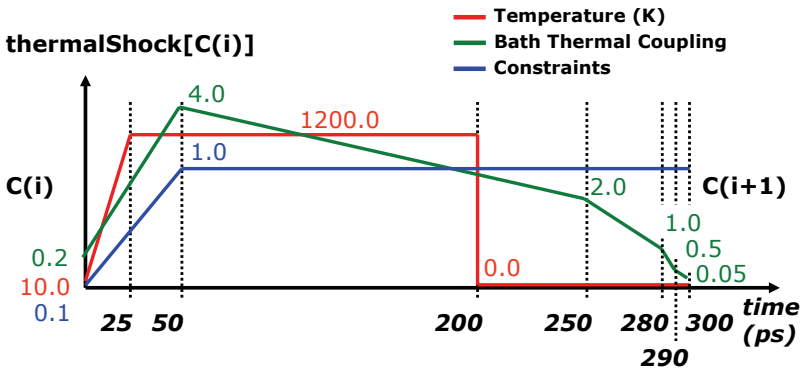


Figure 5.1: Schematic depiction of SA single 300-picoseconds step. Coloured lines, although not scaled- represent the dynamically coupled control parameters.

Descriptors I and II were calculated for the whole chain of conformations although only the second one was extensively used in the present Thesis. Complete information regarding descriptors I and II, Markov matrices, statistics and chain of conformations not included in this chapter can be found in the DVD attached to the back cover.

5.2.1.1 CD5

5.2.1.1.1 Saturation Analysis

This cyclodextrin implements the set of charges by Dr. Beà. The Conformational space of this cyclodextrin has been successfully explored. Saturation Analysis based on Descriptor II fully converges above 256 steps [Fig. 5.2], therefore the total number of conformations found under these conditions is 4 and, with high probability, no other conformations will appear.

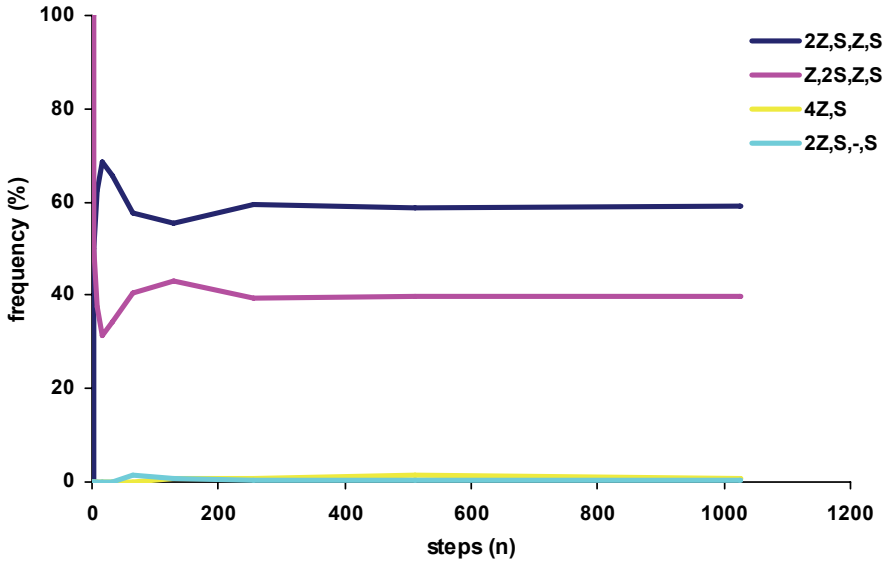


Figure 5.2: Saturation Analysis graph for CD5. 1025 steps. Geometrical approach.

Statistical analysis on the SA ensemble rendered the weight of every conformation along with the average energy and its standard deviation [Fig. 5.3]. In this sense, some remarkable results can be derived:

- Only the first two conformations are of importance, explaining 99% of the system.
- No conformation bearing any **a**-character in Descriptor II was found. This result points to the fact that the high rigidity of CD5 prevents the "inversion" of any glucose in the macroring.
- As a general behaviour in CD5, the total energy grows as the statistical weight decreases. This behaviour is quite common in rigid macromolecules with narrow conformational spaces.

1025 steps conformations		Frequency (N)	(%)	Energy (kcal/mol) average	std.dev.
1	**2Z,S,Z,S**	607	59.22	187.37	5.03
2	**Z,2S,Z,S**	407	39.71	191.20	4.57
3	**4Z,S**	9	0.88	193.33	1.01
4	**2Z,S,-,S**	2	0.20	199.56	2.54
TOTAL (4/4)		**1025**	**100.00**		

Figure 5.3: Data table for CD5. Total number of conformers and their populations were successfully calculated. Average energies and standard deviations are also included.

5.2.1.1.2 Principal Component Analysis

PCA was also applied and, as said in Chapter IV, it suggests that only two principal components (PC's) are necessary to explain almost 99% of the system and conformational groups can be easily detected in well-defined clusters [Fig. 5.4]. In addition, this result was reassuring because it was in accordance with Descriptor II definition and the conformations detected by Saturation Analysis.

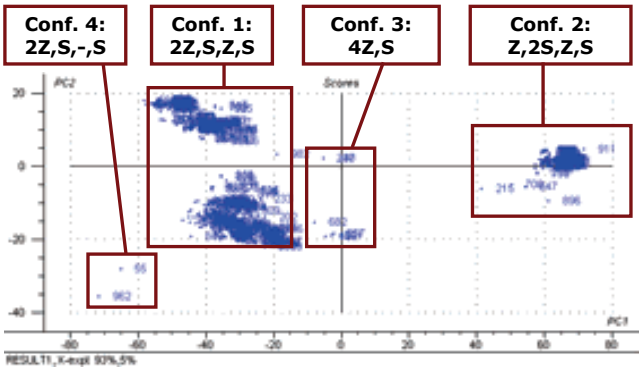


Figure 5.4: PCA graph for CD5. Well-defined conformational clusters clearly separate conformations.

5.2.1.1.3 Markov Analysis

The results obtained clearly pointed that the system not only had been successfully explored but also thermodynamic equilibrium had been reached. In this situation, it was decided to carry out Markov analysis to give an answer to how conformational interchanges occurred [Fig. 5.5].

			State "n+1"			
			2Z,S,Z,S	Z,2S,Z,S	4Z,S	2Z,S,-,S
State "n"	1	2Z,S,Z,S	0.58911	0.40594	0.00330	0.00165
	2	Z,2S,Z,S	0.60197	0.37838	0.01720	0.00246
	3	4Z,S	0.55556	0.44444	0.00000	0.00000
	4	2Z,S,-,S	0.00000	1.00000	0.00000	0.00000

Figure 5.5: Normalized Markov transition matrix for CD5. It contains n^2 boxes, being "n" the total number of different conformations.

Determining Markov transition matrix –when possible- is important because it relates the transition probabilities from a given conformation to any other in the ensemble. I.e. assuming that current state is conformation **1**, [**2Z,S,Z,S**], then there is a probability of 58.91% to remain in this very conformation the next step; 40,59% to change to

conformation **2**, [**Z,2S,Z,S**]; 0.33% to change to conformation **3**, [**4Z,S**]; and 0.16% to change to conformation **4**, [**2Z,S,-,S**]. Reading other conformations from the left column the process is the same.

The equilibrium property, $\rho_{n+1} = \rho_n$, where, $\rho_{n+1} = \{\pi\}^T \rho_n$, is achieved, proving that the system whose Markov matrix was derived is in thermodynamic equilibrium [Eq. 5.1] (see chapter IX in the Annex).

$$\begin{bmatrix} 0.592769 \\ 0.396493 \\ 0.008783 \\ 0.001952 \end{bmatrix} = \begin{bmatrix} 0.589108 & 0.601965 & 0.555555 & 0.000000 \\ 0.405940 & 0.378378 & 0.444444 & 1.000000 \\ 0.003300 & 0.017199 & 0.000000 & 0.000000 \\ 0.001650 & 0.002457 & 0.000000 & 0.000000 \end{bmatrix} \begin{bmatrix} 0.592195 \\ 0.397073 \\ 0.008780 \\ 0.001951 \end{bmatrix}$$

Equation 5.1: Markov product. The dot product of Transition Matrix by weight density vector produces again the weight density vector only when the system is in equilibrium.

This descriptor supplies valuable information; i.e. which conformations can be considered to be *neighbours in the vicinity* according to conformational proximity criteria, low conformational barriers criteria, or both at the same time.

In this sense, PCA diagrams and Markov transitions probabilities altogether can be represented somehow in a *combined graph* that shows the transition paths [Fig. 5.6]. Nevertheless, although this plot is clear, it cannot be regarded as absolutely trustworthy –it is just a *hint*- since the PCA transformation –in this case of CD5- maps a 5-dimensional cyclic space into a 2-dimensional Cartesian space, which in any case guarantees any equivalence in the metrics of the spaces.

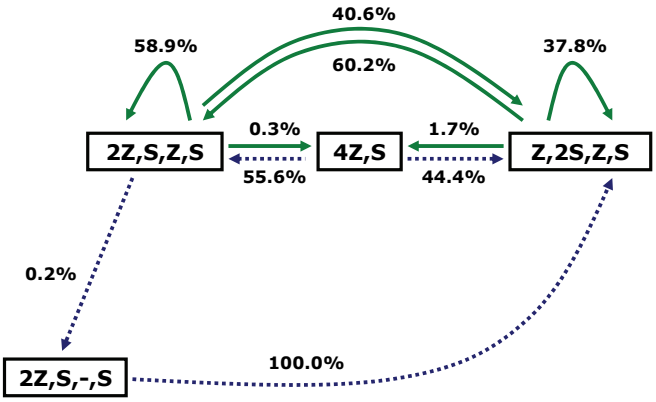


Figure 5.6: Schematic PCA-Markov graph for CD5. Green lines match highly populated conformations while dotted-blue lines link poorly populated ones. As said, shorter distances inter clusters do not necessarily mean higher transition probabilities.

5.2.1.1.4 Results

Combining information and briefly summarizing: the system presents four conformations, although it spends most of its time extensively interchanging between the two most stable [**2Z,S,Z,S**] and [**Z,2S,Z,S**].

5.2.1.2 CD5x

This cyclodextrin is related to CD5 (CD5i) only in part. Both share the same Force Field –parm99- and number of glucoses, nevertheless, the set of charges employed is different; while CD5 implements the set of charges by Dr. Beà, CD5x uses a different one calculated by Dr. Pérez. This particular feature makes the system interesting to be investigated since differences in conformational changes will be –in principle- caused just by this set of charges.

5.2.1.2.1 Saturation Analysis

Again, as in the case of CD5, the Conformational space of this cyclodextrin has been successfully explored. Saturation Analysis based on Descriptor II fully converges above 512 steps [Fig. 5.7], rendering a total number of ten conformations. Likewise, the graph suggests that, with high probability, no other conformations will be found.

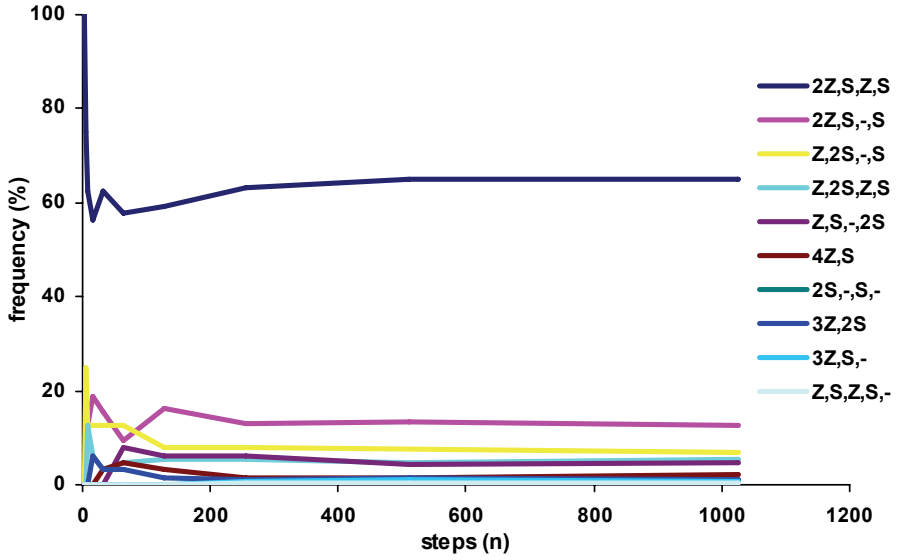


Figure 5.7: Saturation Analysis graph for CD5x. 1025 steps. Geometrical approach.

Further analysis carried out on the SA CD5x ensemble rendered the statistical weights, the average energy and standard deviation [Fig. 5.8]. Some general results can be derived:

- Only the first six conformations are of importance, explaining nearly 96% of the system.
- No conformation bearing any **a**-character in Descriptor II was found. This result –similar to CD5- is the consequence of the high rigidity, which prevents the "inversion" of any glucose in the macroring.
- A different energetic profile was detected in CD5x. Previously, in CD5, the total energy strictly grew as the statistical weight decreased. Nevertheless, this tendency changes here, i.e., allowing conformation 2 to be –energetically speaking- above conformations 3, 4, 5, 6, 8, 9 and 10. This result would appear later in larger CDs and we assumed that it was closely related to entropic effects.

1025 steps	conformations	Frequency (N)	Frequency (%)	Energy (kcal/mol) average	Energy (kcal/mol) std.dev.
1	**2Z,S,Z,S**	667	65.07	87.06	6.83
2	**2Z,S,-,S**	129	12.59	98.18	5.19
3	**Z,2S,-,S**	69	6.73	92.05	7.21
4	**Z,2S,Z,S**	55	5.37	92.74	5.34
5	**Z,S,-,2S**	49	4.78	95.85	4.90
6	**4Z,S**	23	2.24	91.89	3.47
7	**2S,-,S,-**	11	1.07	101.47	7.77
8	**3Z,2S**	10	0.98	93.57	2.55
9	**3Z,S,-**	9	0.88	91.93	4.19
10	**Z,S,Z,S,-**	3	0.29	94.38	6.05
TOTAL (10/10)		**1025**	**100.00**		

Figure 5.8: Data table for CD5x. Total number of conformers and their populations, average energies and standard deviations.

In comparison to CD5, where four conformations were found under similar conditions, CD5x seems to have ten, with only conformations 1, 2, 4 and 6 being common to both of them. The other ones are exclusively detected in CD5x. Assuming that the only difference between CDs is the set of charges employed, then we can conclude that charge selection is a matter of importance regarding SA calculations.

5.2.1.2.2 Principal Component Analysis

PCA showed that, at least, 3 PC's were necessary to explain 93% of the system. This result contrasts with that of CD5, where only 2 PC's explained 99% of the system.

However, the conformational space of CD5x is more complex –10 conformations appearing instead of 4- and therefore three projections –pc1 vs. pc2, pc1 vs. pc3, and pc2 vs. pc3- are required for a complete comprehension.

The first projection –pc1 vs. pc2- [Fig. 5.9] explains 85% of the system and shows good separations for conformations 1, 2, 3, 6, 7 and 9. Nevertheless, conformations 1, 2 and 3 spread in a continuum area and conformations 4, 5, 8 and 10 seem to be partially overlapped. Seeking for better resolution for some of these groups other projections are studied.

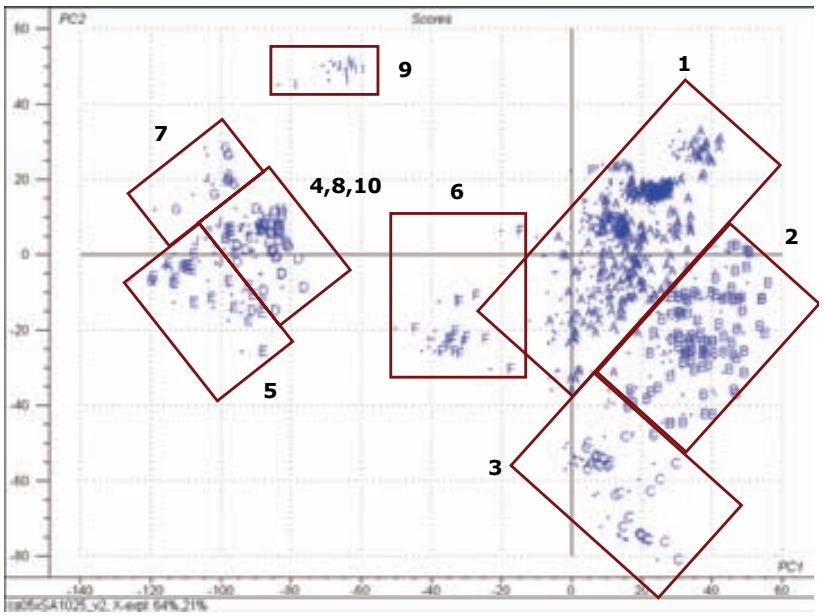


Figure 5.9: pc1 vs. pc2 graph for CD5x.

The second projection –pc1 vs. pc3- [Fig. 5.10] explains 72% of the system and establishes differences between conformations 6, 7, 9 and 10. It is interesting to remark that this projection clearly separates conformation 10 out of the group in [Fig. 5.9] and also 7 appears in a lower layer below 4, 5, and 8. Again, conformations 1, 2, and 3 appear closely distributed in 2 groups; however, this is not the best point of view and at this time it is difficult to draw a line that delineates the areas of influence of any particular conformation. The group of conformations 4, 5, and 8 behaves in a similar way.

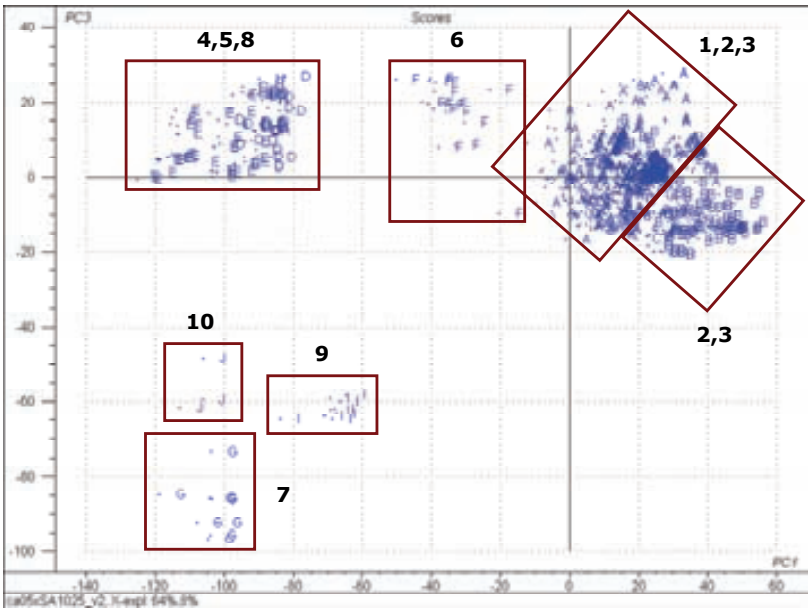


Figure 5.10: pc1 vs. pc3 graph for CD5x.

The last projection studied –pc2 vs. pc3- [Fig. 5.11] explains 29% of the system and is particularly useful in separating the clusters of conformations 3, 7, 9 and 10. Finally, one of the most important conformations in CD5x, namely 3, has been successfully delimited out of the continuum group involving 1, and 2. Nevertheless this projection includes a large cluster including overlapped conformations 1, 2, 4, 5, 6 and 8.

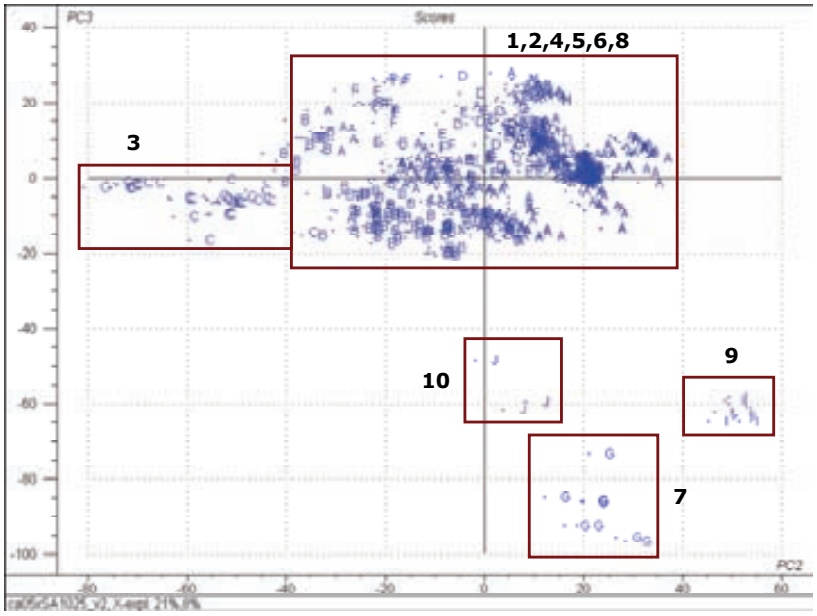


Figure 5.11: pc2 vs. pc3 graph for CD5x.

The fact that conformations 1, 2, and 3 clot together in a single continuum area seems to point that all of them are close –conformationally speaking-.

Probably, the most important conclusion is that, as flexibility grows, conformations avoid isolated areas and appear in groups that continually interconvert, not allowing easy differentiation.

5.2.1.2.3 Markov Analysis

Thermodynamic equilibrium was achieved (Markov dot product is included in the DVD) and Markov analysis was applied to the SA ensemble; the transition probabilities were successfully determined for all conformations [Fig. 5.12].

	1	2	3	4	5	6	7	8	9	10
1	0.64414	0.13814	0.06607	0.04955	0.04204	0.01952	0.01201	0.01502	0.00901	0.00450
2	0.67442	0.12403	0.07752	0.04651	0.05426	0.01550	0.00775	0.00000	0.00000	0.00000
3	0.63768	0.13043	0.08696	0.10145	0.01449	0.02899	0.00000	0.00000	0.00000	0.00000
4	0.61818	0.07273	0.03636	0.07273	0.10909	0.03636	0.01818	0.00000	0.03636	0.00000
5	0.55102	0.12245	0.10204	0.08163	0.04082	0.06122	0.02041	0.00000	0.02041	0.00000
6	0.73913	0.04348	0.04348	0.04348	0.08696	0.04348	0.00000	0.00000	0.00000	0.00000
7	0.90909	0.00000	0.00000	0.00000	0.09091	0.00000	0.00000	0.00000	0.00000	0.00000
8	0.80000	0.10000	0.00000	0.00000	0.10000	0.00000	0.00000	0.00000	0.00000	0.00000
9	0.88889	0.00000	0.11111	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	0.66667	0.00000	0.00000	0.00000	0.33333	0.00000	0.00000	0.00000	0.00000	0.00000

Figure 5.12: Normalized Markov transition matrix for CD5x.

5.2.1.2.4 Results

Summarising: The system has 10 conformations, although most of the time it is in conformation **1**, [**2Z,S,Z,S**]. The point is that, under SA process, all of them seem to have a high probability to interconvert in the next step into conformation **1** (check first column), and conformation **1** has a high probability of remaining in **1** (check first row) instead of changing to any other. Comparing PCA with Markov analysis we also realise that conformations involving the higher weights and transition probabilities; **1**, **2**, and **3**, cover a well-defined continuum closed area in conformational space.

This statistical behaviour along with energy results confirms conformation **1** as the principal one.

5.2.1.3 CD6 (α-CD)

5.2.1.3.1 Saturation Analysis

This cyclodextrin implements again the set of charges by Dr. Beà. The Conformational space of CD6 has been successfully explored. Saturation Analysis based on Descriptor II approximately converges above 512 steps [Fig. 5.13], rendering a total number of 18 conformations. The slopes of the lines in the graphic slowly tend to zero suggesting that, with high probability, no other important conformations will be found.

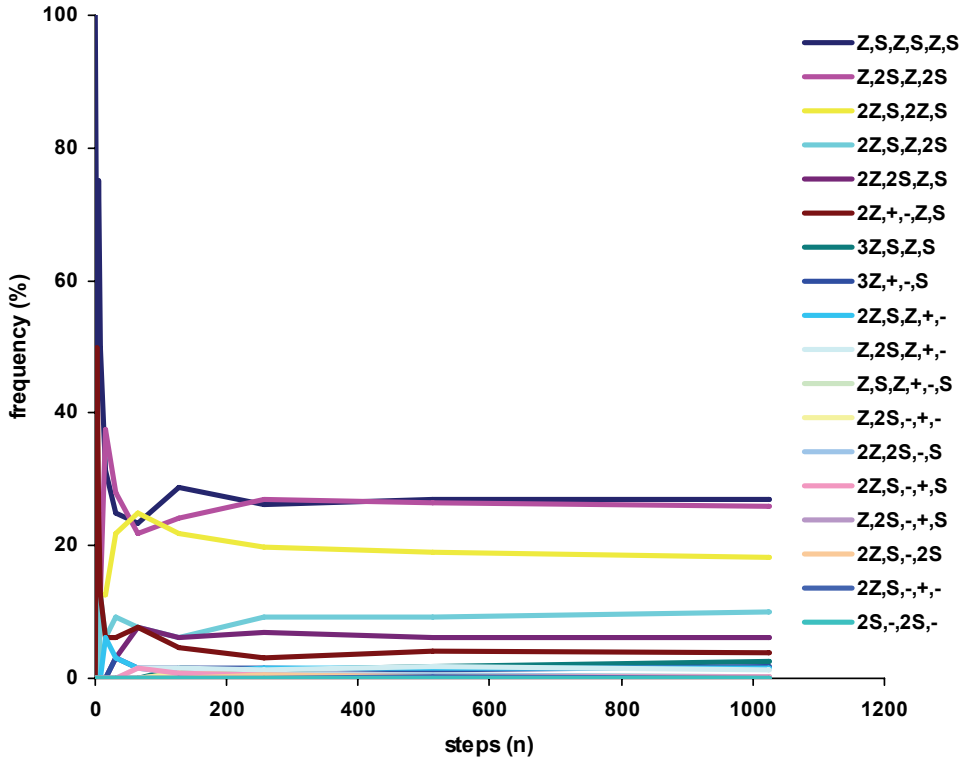


Figure 5.13: Saturation Analysis graph for CD6. 1025 steps. Geometrical approach.

Further analysis carried out on the SA CD6 ensemble rendered the statistical weights, the average energy and the standard deviation [Fig. 5.8]:

- Only the first six conformations are of importance, explaining nearly 90% of the system. Anyway, the first four are the most highly weighted ones in the ensemble.
- No conformation bearing any **a**-character in Descriptor II was found although this result is however contradicted by Descriptor I, that includes 7 instances of

conformation [**4s,+,a**] (check DVD). When D-II was defined some borders were shifted and this result directly points to this fact. Anyway, previous analysis for CD5 and CD5x did not fall in this *inconsistency* between D-I and D-II that is ultimately the consequence of growing flexibility in the macroring:

As the number of glucoses grows the system is close to glucose inversion.

- The energetic profile is in someway *relaxed*. The expected behaviour was that the total energy should strictly grow as the statistical weight decreases, but here, this tendency disappears. On the one hand, the most populated conformation is higher in energy than the second one, while the other conformations do not follow any particular tendency in this point. On the other hand, all conformations display comparable energetic values and standard deviations. This hint suggests that energetic distributions clearly overlap, not allowing energetic criteria to be of use when filtering/analysing conformations:

  As the number of glucoses grows, the flexibility grows.

  Structurally different conformations become closer in energy.

1025 steps conformations		Frequency (N)	(%)	Energy (kcal/mol) average	std.dev.
1	**Z,S,Z,S,Z,S**	277	27.02	216.31	14.81
2	**Z,2S,Z,2S**	265	25.85	202.42	17.04
3	**2Z,S,2Z,S**	187	18.24	220.29	16.71
4	**2Z,S,Z,2S**	103	10.05	220.55	16.03
5	**2Z,2S,Z,S**	63	6.15	224.13	14.46
6	**2Z,+,-,Z,S**	40	3.90	229.71	13.67
7	**3Z,S,Z,S**	27	2.63	218.72	17.90
8	**3Z,+,-,S**	18	1.76	239.08	10.64
9	**2Z,S,Z,+,-**	17	1.66	227.08	13.30
10	**Z,2S,Z,+,-**	13	1.27	230.66	14.00
11	**Z,S,Z,+,-,S**	3	0.29	229.49	6.37
12	**Z,2S,-,+,-**	3	0.29	230.66	2.25
13	**2Z,2S,-,S**	3	0.29	232.55	26.55
14	**2Z,S,-,+,S**	2	0.20	227.99	12.90
15	**Z,2S,-,+,S**	1	0.10	264.16	(-)
16	**2Z,S,-,2S**	1	0.10	225.09	(-)
17	**2Z,S,-,+,-**	1	0.10	248.80	(-)
18	**2S,-,2S,-**	1	0.10	218.67	(-)
TOTAL (18/18)		**1025**	**100.00**		

Figure 5.14: Data table for CD6. Total number of conformers and their populations, average energies and standard deviations.

5.2.1.3.2 Principal Component Analysis

In this case, PCA showed that only 2 PC's were necessary to explain 92% of the system. This result is interesting because despite having 18 conformations –and

therefore a rather complex conformational space- just a single projection –pc1 vs. pc2- is enough for a full description of the system [Fig. 5.15].

In general, the group of conformations falls in sets of rather well-separated areas, the most important ones lying in the middle of the plot. The clusters forming the central "U" gather the main –most populated- conformations; **1**, **2**, **3**, **4** and **5**.

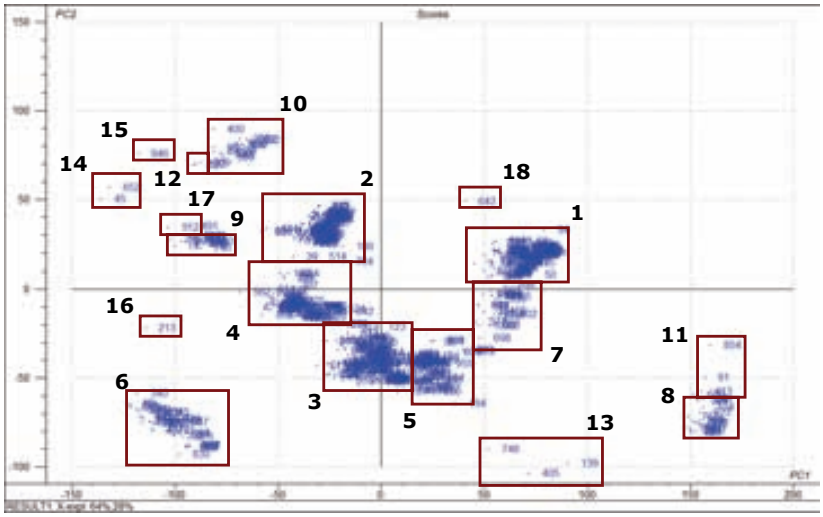


Figure 5.15: PCA graph for CD6. Well-defined conformational clusters clearly separate conformations.

5.2.1.3.3 Markov Analysis

The CD6 ensemble reached thermodynamic equilibrium (Full Markov dot product is included in the DVD) and Markov analysis was applied; the transition probabilities were successfully determined for all the conformations [Fig. 5.16].

	1	2	3	4	5	6	7	8	9	10
1	0.25271	0.27437	0.21300	0.09025	0.06859	0.04693	0.01444	0.01083	0.00361	0.01083
2	0.25283	0.20755	0.21887	0.11698	0.07170	0.03396	0.03019	0.01887	0.02264	0.00755
3	0.29032	0.29032	0.13978	0.09677	0.04839	0.03763	0.04301	0.01075	0.02688	0.00000
4	0.24272	0.26214	0.18447	0.05825	0.06796	0.05825	0.03883	0.04854	0.00000	0.03883
5	0.31746	0.25397	0.17460	0.09524	0.01587	0.04762	0.03175	0.03175	0.00000	0.01587
6	0.40000	0.25000	0.10000	0.10000	0.07500	0.00000	0.00000	0.02500	0.02500	0.02500
7	0.22222	0.29630	0.14815	0.14815	0.03704	0.03704	0.03704	0.00000	0.00000	0.03704
8	0.27778	0.22222	0.16667	0.11111	0.11111	0.00000	0.00000	0.00000	0.05556	0.05556
9	0.35294	0.35294	0.05882	0.17647	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	0.23077	0.15385	0.15385	0.30769	0.00000	0.00000	0.00000	0.00000	0.15385	0.00000

Figure 5.16: Normalized Markov transition sub-matrix for CD6. Only the first ten conformations – those representing 98% of the system- have been included. Full transition matrix can be found in the DVD.

5.2.1.3.4 Results

In summary: The system has –with high probability- only 18 conformations although most of the time, it is in one of the first four conformations; **1**, [**Z,S,Z,S,Z,S**]; **2**, [**Z,2S,Z,2S**]; **3**, [**2Z,S,2Z,S**]; and **4**, [**2Z,S,Z,2S**], being the first two the preferred ones. Under detailed examination of the transition matrix –4x4 submatrix- and statistical weights –first four values-, it is easy to realise that the 3 –perhaps 4- most populated conformations interconvert, thus representing the primary path in conformational interchanges. Hence, this reduced group of conformations covers a well-defined continuum close area in conformational space and acts entirely as an attractor, driving the behaviour of the system towards this well. This analysis confirms conformations **1**, **2**, **3**, and **4** as the principal ones.

5.2.1.4 CD7 (β-CD)

5.2.1.4.1 Saturation Analysis

CD7 implements again the set of charges by Dr. Beà. Now, as the Saturation Analysis seems to suggest, the research is entering the flexible group of CDs. The Conformational space of CD7 has been –in principle- satisfactorily explored: The slopes in Descriptor II saturation graphic slightly tends to zero although they never seem to really converge within the 1025 steps [Fig. 5.17]. However, a total number of 66 conformations were found. This result suggests that, probably, populated conformations will maintain their status and only unimportant ones would be found if longer calculations were done.

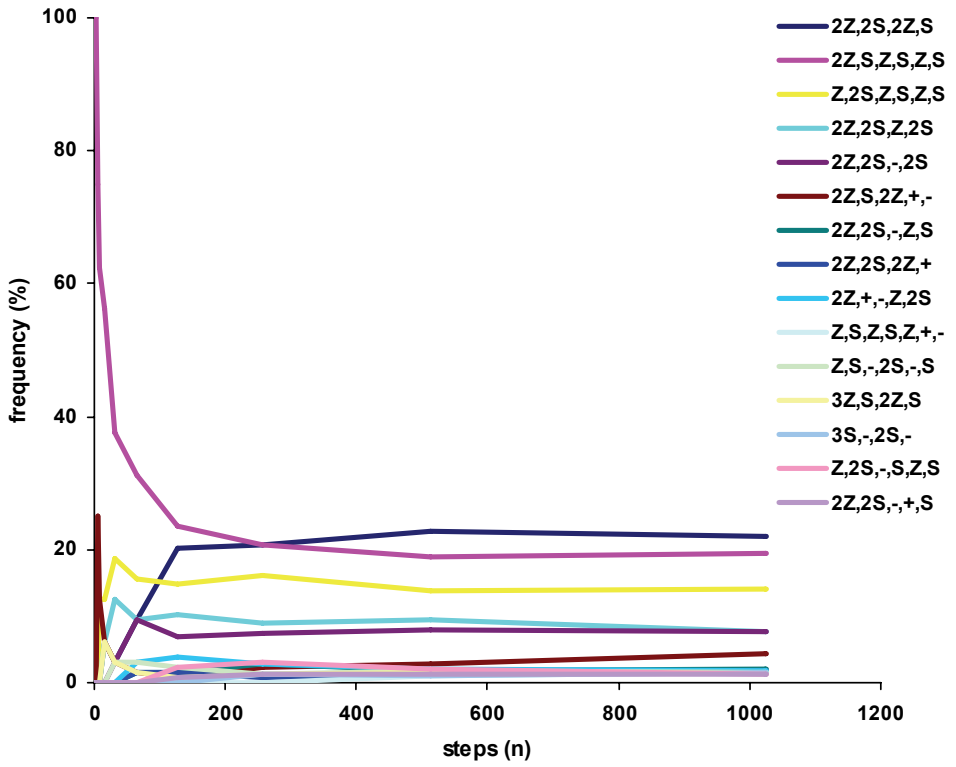


Figure 5.17: Saturation Analysis graph for CD7. 1025 steps. Geometrical approach. Only the first fifteen conformations –the most important ones, representing 89% of the system- have been analysed.

As was done in previous CDs, statistical analysis was carried out on the SA CD7 ensemble rendering the statistical weights, average energy and standard deviation [Fig. 5.18]:

1025 steps conformations		Frequency (N)	(%)	Energy (kcal/mol) average	std.dev.
1	**2Z,2S,2Z,S**	226	22.05	245.87	22.96
2	**2Z,S,Z,S,Z,S**	199	19.41	258.92	25.60
3	**Z,2S,Z,S,Z,S**	144	14.05	263.98	23.99
4	**2Z,2S,Z,2S**	79	7.71	224.17	27.13
5	**2Z,2S,-,2S**	78	7.61	239.95	26.39
6	**2Z,S,2Z,+,-**	45	4.39	248.79	22.48
7	**2Z,2S,-,Z,S**	21	2.05	269.20	21.68
8	**2Z,2S,2Z,+**	19	1.85	273.49	15.67
9	**2Z,+,-,Z,2S**	18	1.76	243.30	21.19
10	**Z,S,Z,S,Z,+,-**	16	1.56	268.10	26.95
11	**Z,S,-,2S,-,S**	15	1.46	272.37	24.44
12	**3Z,S,2Z,S**	15	1.46	234.27	23.46
13	**3S,-,2S,-**	15	1.46	261.05	26.12
14	**Z,2S,-,S,Z,S**	14	1.37	258.21	25.81
15	**2Z,2S,-,+,S**	12	1.17	273.03	18.62
TOTAL (15/66)		**916**	**89.37**		

Figure 5.18: Data table for CD7. The first 15 conformers –out of a total number of 66- and their populations, average energies and standard deviations have been extracted. The statistical weight of this group of conformations is nearly 90%.

- Only the first five conformations are of importance, explaining nearly 70% of the system. However, the first three are the most highly weighted ones in the ensemble.
- 22 conformations –one in three- bearing **a**-characters in Descriptor II –and other in Descriptor I- were found [Fig. 5.19] (check DVD) weighting nearly 3.9%. This cyclodextrin is clearly more flexible than those previously explained and inversion of glucose rings within the macroring is detected as the consequence of it.

N	conformation	N	conformation
6	2Z,a,-,2Z,+	1	Z,S,-,a,-,2S
4	3Z,2S,-,a	1	Z,+,a,2S,-,S
3	2Z,S,2Z,a,-	1	Z,2S,-,a,-,S
3	2Z,3S,-,a	1	3Z,S,a,2S
3	2S,a,+,2S,-	1	3Z,+,S,-,a
2	Z,S,Z,S,a,+,S	1	3Z,2S,a,S
2	2Z,S,a,+,Z,S	1	2Z,S,-,a,-,S
2	2Z,a,-,Z,2S	1	2Z,a,-,S,-,S
2	2Z,2S,Z,a,-	1	2Z,+,a,3S
1	Z,S,Z,S,Z,a,-	1	2Z,2S,a,+,S
1	Z,S,Z,S,-,a,-	1	2Z,2S,-,a,-

Figure 5.19: Set of CD7 conformations where glucose inversion was found. Columns on the left include the number of times they were found.

- Apparently, there is no relationship between statistical weight and energetic stability; therefore, highly populated conformations are not necessarily more stable than others less populated –further information will be given in Hydrogen Bond section 5.3.4-. Besides, all the conformations fall in a range of comparable energetic values and standard deviations, and that enables overlapping between energetic distributions. Again, it is derived that:
  As the number of glucoses grows, the flexibility grows.
  Conformations structurally different become closer in energy.

5.2.1.4.2 Principal Component Analysis

PCA was applied concluding that 3 PC's –pc1, pc2 and pc3- were necessary to explain 91% of the system, this time requiring a 3-dimensional plot. Besides, it is observed that the complexity of the conformational space rises up as the number of glucoses grows. Two different points of view of the 3D plot have been selected; however, the entire

group of clusters cannot be seen as a clearly defined one all at the same time, because when resolution is good for some of them, then the others are partially eclipsed.

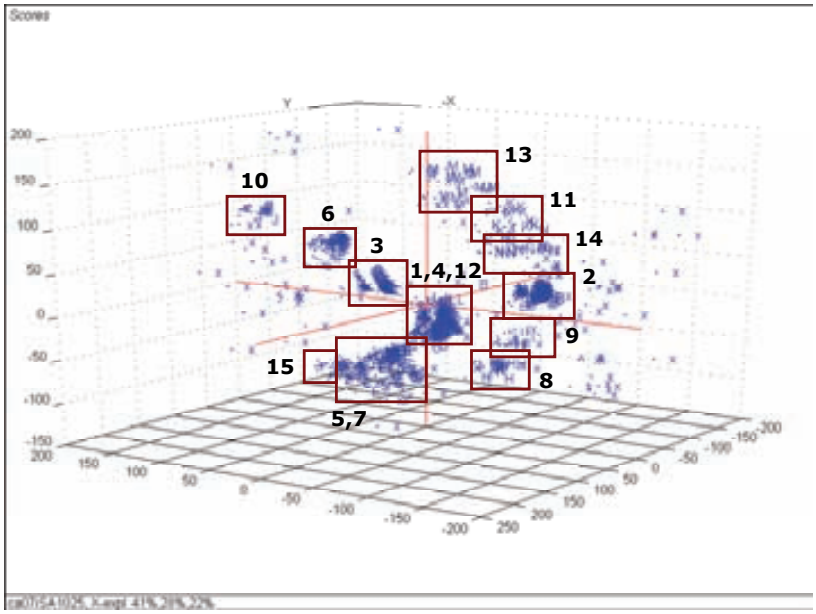


Figure 5.20: 3D graphic including pc1 vs. pc2 vs. pc3 for CD7. Point of view [–x, y, z].

An alternative point of view of the 3D plot [Fig. 5.21] is also included to make clear that the effect of rotating the coordinate frame improves the resolution of some groups at the expense of others. I.e., while conformation **8** is well defined in [Fig. 5.20], now in [Fig. 5.21] it is partially overlapped by **12**.

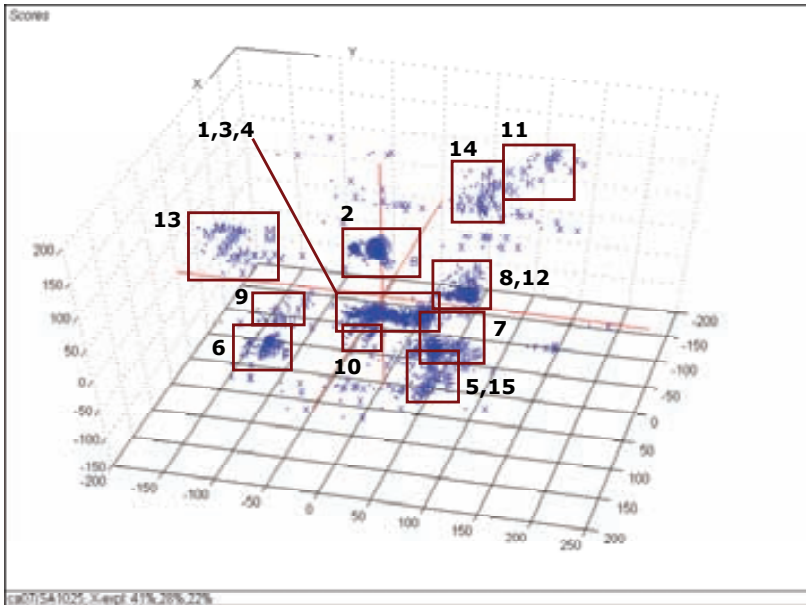


Figure 5.21: 3D graphic including pc1 vs. pc2 vs. pc3 for CD7. Point of view [x, y, z].

Here, a new scenario is faced: Detailed examination of the two graphics [Fig. 5.20 and 5.21] suggests that conformations should be divided in two types:

- The highly populated ones, representing the main conformations: A = 1, B = 2, and so on until O = 15.
- The poorly populated ones, representing unimportant conformations: x = 16, 17, 18... 66.

In general, the first ones –those most populated- are represented by dense clusters of tightly packed sets of conformations. They usually fall near the origin of coordinates in well-defined groups with reasonable resolution, although sometimes the solution of continuity involving highly populated areas is not clear –i.e. conformations 1, 3 and 4 in [Fig. 5.21]-. The second ones –those with low ratios- spread all over conformational space as an amorphous foggy cloud, not really defining any particular cluster but a *conformational noise*.

This situation is new and contrasts with the behaviour observed in CD5, CD5x and CD6, where only conformations in the type of the first group were detected.

5.2.1.4.3 Markov Analysis

CD7 ensemble seems to have reached –reasonably speaking- thermodynamic equilibrium as was said at the beginning of this section and, as in the previous cases, Markov analysis was applied and the transition probabilities were successfully determined for all the conformations [Fig. 5.22] (Full Markov dot product is included in the DVD).

	1	2	3	4	5	6	7	8	9	10
1	0.24889	0.20444	0.14222	0.07111	0.06667	0.05778	0.01333	0.02222	0.01333	0.00889
2	0.22111	0.23116	0.13065	0.06533	0.06533	0.03015	0.02513	0.01005	0.01508	0.02513
3	0.18750	0.18056	0.13194	0.06944	0.06944	0.06250	0.03472	0.00694	0.02083	0.00694
4	0.21519	0.10127	0.29114	0.07595	0.03797	0.02532	0.00000	0.03797	0.02532	0.01266
5	0.24359	0.20513	0.10256	0.05128	0.08974	0.05128	0.01282	0.03846	0.01282	0.00000
6	0.22222	0.13333	0.17778	0.20000	0.04444	0.02222	0.00000	0.02222	0.04444	0.02222
7	0.19048	0.23810	0.04762	0.09524	0.04762	0.04762	0.00000	0.00000	0.00000	0.04762
8	0.21053	0.05263	0.21053	0.05263	0.21053	0.05263	0.00000	0.00000	0.05263	0.00000
9	0.16667	0.33333	0.11111	0.05556	0.05556	0.05556	0.00000	0.05556	0.00000	0.00000
10	0.25000	0.18750	0.06250	0.06250	0.25000	0.00000	0.00000	0.06250	0.06250	0.00000

Figure 5.22: Normalized Markov transition sub-matrix for CD7. Only the first ten conformations – those representing 82% of the system- have been included. Full transition matrix can be found in the CD.

5.2.1.4.4 Results

In summary: The system has –with high probability- not more than 7 important conformations although most of the time, it is in one of the first three ones; **1**, [**2Z,2S,2Z,S**]; **2**, [**2Z,S,Z,S,Z,S**]; and **3**, [**Z,2S,Z,S,Z,S**].

Transition matrix gives also a hint regarding the new conformational behaviour detected in CD7. As mentioned, two types were defined; the *main conformations* and the *conformational noise*. Detailed examination of the 10x10 Markov submatrix shows several boxes in the lower triangle filled with probabilities other than zero. Now, this result is interesting since it is becoming an ordinary phenomenon in larger CDs. Examining Markov matrices and submatrices in CD5, CD5x and CD6, the boxes in the lower triangles usually included probabilities of zero, or close to zero; meaning that:

- Highly populated conformations tend to interconvert into highly populated conformations.
- Highly populated conformations rarely interconvert into poorly populated conformations.
- Poorly populated conformations tend to convert into highly populated conformations.
- Poorly populated conformations almost never convert into poorly populated conformations.

This situation has changed when considering CD7. The appearance of values other than zero in the lower triangle means that:

- Highly populated conformations not only interconvert into highly populated ones, but also into others less populated.
- Lowly populated conformations not only interconvert into highly populated conformations, but they sometimes continue exploring other underpopulated ones.

This analysis confirms the first five conformations as the principal ones, including also the conformational noise behaviour.

5.2.1.5 CD8 (γ-CD)

5.2.1.5.1 Saturation Analysis

CD8 implements the set of charges by Dr. Beà. The situation is similar to that of CD7; however, the Saturation Analysis for CD8 shows an even more flexible macromolecule. The slopes in Descriptor II graphic –asymptotically approximating to zero- prove that the conformational space has been satisfactorily explored [Fig. 5.23] although a higher number of steps –not less than 4096- would be advisable since 1024 steps were not enough to ensure full convergence. Nevertheless, the results obtained for CD8 can be regarded as trustworthy: A total number of 135 conformations were found, the first 16 ones weighting nearly 77%, this meaning that, probably, highly populated conformations will remain in their ranking and only unimportant ones –those previously called *conformational noise*- will be found if longer calculations are done.

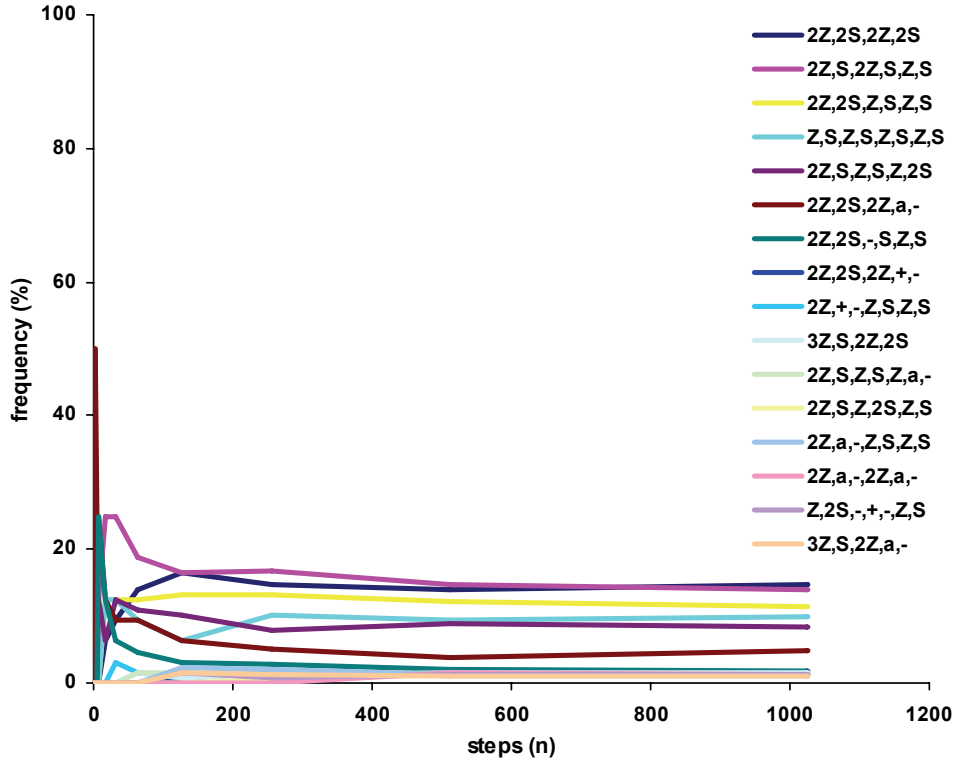


Figure 5.23: Saturation Analysis graph for CD8. 1025 steps. Geometrical approach. Only the first sixteen conformations –the most important ones, representing 77% of the system- have been analysed.

The statistical analysis carried out on the SA CD8 ensemble rendered the statistical weights, average energies and standard deviations shown in [Fig. 5.24]:

1025 steps conformations		Frequency (N)	(%)	Energy (kcal/mol) average	std.dev.
1	**2Z,2S,2Z,2S**	150	14.63	248.28	20.37
2	**2Z,S,2Z,S,Z,S**	144	14.05	248.59	22.18
3	**2Z,2S,Z,S,Z,S**	116	11.32	263.19	23.66
4	**Z,S,Z,S,Z,S,Z,S**	102	9.95	287.89	24.02
5	**2Z,S,Z,S,Z,2S**	86	8.39	276.54	20.85
6	**2Z,2S,2Z,a,-**	49	4.78	291.32	24.64
7	**2Z,2S,-,S,Z,S**	17	1.66	280.06	30.87
8	**2Z,2S,2Z,+,-**	17	1.66	280.13	18.41
9	**2Z,+,-,Z,S,Z,S**	16	1.56	284.76	25.43
10	**3Z,S,2Z,2S**	15	1.46	277.90	17.13
11	**2Z,S,Z,S,Z,a,-**	14	1.37	274.53	30.04
12	**2Z,S,Z,2S,Z,S**	13	1.27	260.56	20.58
13	**2Z,a,-,Z,S,Z,S**	13	1.27	290.02	20.30
14	**2Z,a,-,2Z,a,-**	13	1.27	301.25	25.54
15	**Z,2S,-,+,-,Z,S**	12	1.17	297.52	24.68
16	**3Z,S,2Z,a,-**	10	0.98	275.83	16.45
TOTAL (16/135)		**787**	**76.78**		

Figure 5.24: Data table for CD8. The first 16 conformers –out of a total number of 135- and their populations, average energies and standard deviations have been extracted. The statistical weight of this group of conformations is nearly 77%.

- Only the first six conformations are of importance, explaining nearly 63% of the system. However, the first three are the most highly weighted ones in the ensemble.
- 59 conformations –almost one in two- bearing **a**-characters in Descriptor II –and other in Descriptor I- were found [Fig. 5.25] (check DVD) weighting 18.8%. CD8 is clearly more flexible than those previously studied –including CD7- and inversion of glucose rings is detected as a consequence of this.

N	conformation	N	conformation	N	conformation	N	conformation
49	2Z,2S,2Z,a,-	2	3Z,a,-,2Z,S	1	Z,S,-,2S,a,2S	1	2Z,+,Z,-,Z,a,-
14	2Z,S,Z,S,Z,a,-	2	3Z,3S,-,a	1	Z,a,-,S,-,S,+,-	1	2Z,S,a,+,Z,2S
13	2Z,a,-,Z,S,Z,S	2	3Z,2S,Z,-,a	1	Z,a,+,3S,-,S	1	2Z,S,-,2S,-,a
13	2Z,a,-,2Z,a,-	2	3Z,2S,a,2S	1	Z,a,+,2S,-,a,-	1	2Z,S,-,+,2-,a
10	3Z,S,2Z,a,-	2	2Z,a,-,Z,-,2S	1	Z,2S,-,Z,+,a,S	1	2Z,a,Z,-,S,a,-
8	4Z,2S,-,a	2	2Z,3S,-,a,-	1	Z,2S,a,+,S,Z,S	1	2Z,a,-,Z,+,-,S
8	2Z,S,Z,S,-,a,-	2	2Z,2S,-,Z,a,-	1	Z,-,2S,-,a,-,S	1	2Z,a,-,S,Z,2S
7	2Z,S,Z,a,-,Z,S	1	Z,S,Z,S,-,S,a,-	1	Z,2S,a,+,-,a,-	1	2Z,a,-,S,-,2S
4	3Z,S,Z,S,-,a	1	Z,S,Z,-,S,-,a,-	1	Z,2S,a,3S,-	1	2Z,a,-,S,-,+,-
4	2Z,a,2-,S,Z,S	1	Z,S,Z,a,-,Z,a,-	1	Z,-,2S,a,3S	1	2Z,+,a,2S,Z,S
3	Z,S,Z,a,-,Z,+,-	1	Z,S,Z,a,2-,a,-	1	Z,2S,-,+,-,a,-	1	2Z,2S,Z,-,a,-
3	2Z,S,Z,2S,-,a	1	Z,S,Z,a,2-,+,-	1	4Z,a,2-,+	1	2Z,2S,-,S,a,-
2	Z,2S,-,S,Z,a,-	1	Z,S,-,+,Z,-,a,-	1	3Z,S,a,3S	1	2Z,2S,-,a,-,S
2	Z,2S,-,a,-,Z,S	1	Z,S,Z,-,2S,a,+	1	3Z,+,2Z,a,-	1	2Z,2S,a,3S
2	3Z,a,-,Z,-,+	1	Z,S,a,+,S,Z,+,-	1	3S,a,3S,a		

Figure 5.25: Set of CD8 conformations where glucose inversion was found. Columns on the left include the number of times they were found.

- Furthermore, the increasing flexibility produced by the enlargement of the ring allows a new phenomenon not yet observed in the previous CDs. CD8 is the first CD where more than one glucose inversion is permitted within the same macrocyclic ring [Fig. 5.26]. Besides, this is not just a rarity since conformation **14** is one of them.

N	conformation
13	2Z,a,-,2Z,a,-
1	Z,S,Z,a,-,Z,a,-
1	Z,S,Z,a,2-,a,-
1	Z,a,+,2S,-,a,-
1	Z,2S,a,+,-,a,-
1	3S,a,3S,a
1	2Z,a,Z,-,S,a,-

Figure 5.26: Set of CD8 conformations where more than one glucose inversion was found. Columns on the left include the number of times they were found.

- Again, as was said in previous CDs, the increasing flexibility induces that highly populated conformations are not necessarily more stable than others less populated. This fact, along with all the conformations falling in a range of comparable energetic values and standard deviations that enables overlapping between energetic distributions, proves that a*s flexibility grows, conformations structurally different become closer in energy.*

5.2.1.5.2 Principal Component Analysis

PCA concluded that 3 PC's –pc1, pc2 and pc3- were necessary to explain 86% of the system. Again, a single 3D-plot was used instead of a collection of several 2D-plots seeking for the best way of depicting the Conformational space. As in the case of CD7; when the flexibility grows the complexity of the Conformational space is increased, and therefore, higher dimensions are required to give a reasonable explanation of the system.

Two different points of view of the 3D plot have been selected to offer the best description of the system; however, the entire group of clusters cannot be seen clearly defined all at the same time.

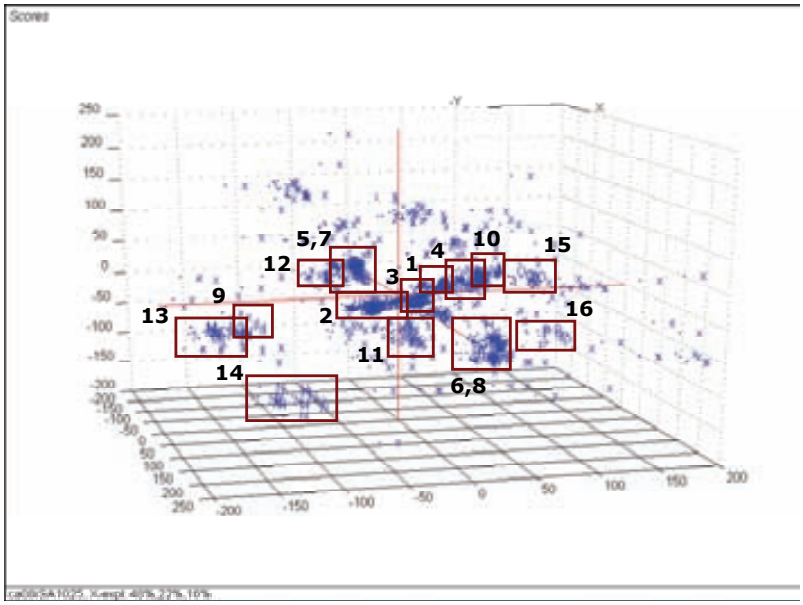


Figure 5.27: 3D graphic including pc1 vs. pc2 vs. pc3 for CD8. Point of view [x, –y, z].

As said in the case of CD7, an alternative point of view of the 3D plot [Fig. 5.27] is also included to make clear that the effect of rotating the coordinate frame improves the resolution of some groups at the expense of others.

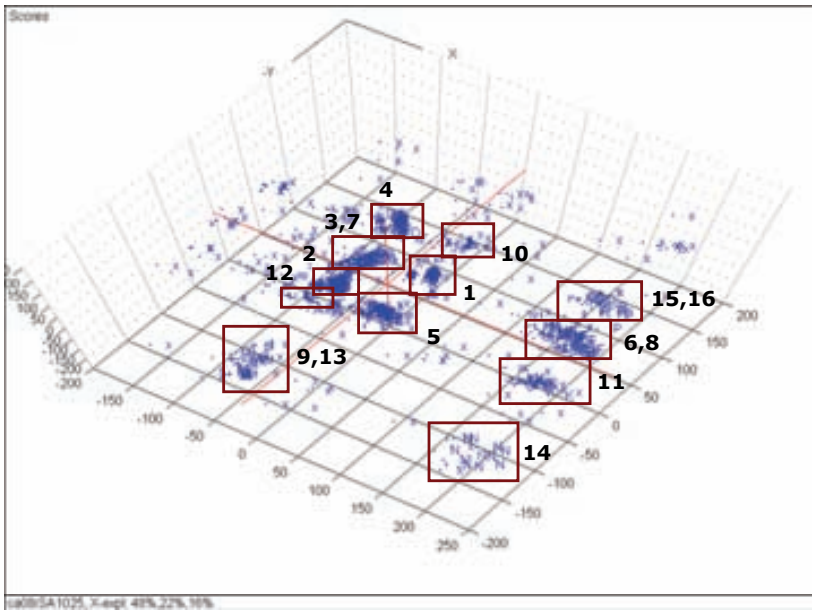


Figure 5.28: 3D graphic including pc1 vs. pc2 vs. pc3 for CD8. Point of view [x, –y, z].

As can be seen in both graphics [Fig. 5.27 and 5.28], the situation in CD8 is similar to that explained in CD7. Two types of conformational behaviour are observed:

- The main conformations, represented by dense clusters of well-defined groups with reasonable resolution: A = 1, B = 2, and so on until P = 16.
- The conformational noise, not really defining any particular cluster: x = 17, 18, 19... 135.

As usual, the first ones exhibit a tendency to be the most populated while the second ones predominate among the low populated.

This situation –common to CD7 and CD8- contrasts with that observed in CD5, CD5x and CD6, where only conformations in the type of the first group were detected: The first important qualitative and quantitative change in the behaviour has been detected.

5.2.1.5.3 Markov Analysis

Saturation Analysis suggests that the CD8 ensemble has almost reached thermodynamic equilibrium. This means that Markov transition matrix can be reasonably obtained from the chain of descriptors [Fig. 5.29] (Full Markov dot product is included in the DVD).

	1	2	3	4	5	6	7	8	9	10
1	0.14667	0.11333	0.10667	0.08667	0.10000	0.06000	0.02000	0.01333	0.01333	0.02667
2	0.14583	0.11806	0.10417	0.15278	0.07639	0.04861	0.02778	0.02083	0.02083	0.02083
3	0.20690	0.17241	0.12931	0.11207	0.07759	0.05172	0.01724	0.00862	0.01724	0.00000
4	0.13725	0.18627	0.12745	0.09804	0.06863	0.07843	0.00980	0.01961	0.01961	0.00980
5	0.20930	0.12791	0.05814	0.02326	0.05814	0.02326	0.01163	0.02326	0.01163	0.01163
6	0.20408	0.20408	0.14286	0.08163	0.08163	0.06122	0.04082	0.00000	0.00000	0.00000
7	0.11765	0.11765	0.05882	0.00000	0.17647	0.11765	0.05882	0.00000	0.00000	0.00000
8	0.06250	0.18750	0.12500	0.25000	0.06250	0.00000	0.00000	0.00000	0.00000	0.00000
9	0.18750	0.12500	0.06250	0.25000	0.12500	0.00000	0.00000	0.00000	0.06250	0.00000
10	0.06667	0.13333	0.00000	0.13333	0.00000	0.00000	0.00000	0.13333	0.00000	0.06667

Figure 5.29: Normalized Markov transition sub-matrix for CD8. Only the first ten conformations – those representing 69.5% of the system- have been included. Full transition matrix can be found in the DVD.

Nevertheless, as the number of glucoses in the CDs grows, also the total number of conformers found by saturation analysis increases and then a new problem is faced: determining Markov transition matrix on the grounds of a single chain of 1025 steps is gradually –statistically speaking- not representative. This problem was also detected in CD7 and is studied in detail here in CD8.

As can be seen, the number of transitions to be computed to obtain the Markov matrix grows as the square of the number of conformations. Therefore, when the number of

entries in the matrix approaches –or exceeds- the number of steps, the statistics regarding transitions involving poorly populated conformations become absolutely untrustworthy [Fig. 5.30]. That is the reason for including only Markov sub-matrices, especially for CDs above CD6.

		CD5	CD5x	CD6	CD7	CD8
Number of steps	**s**	1025	1025	1025	1025	1025
Number of conformers	**c**	4	10	18	66	135
Markov matrix size	**$z = c^2$**	16	100	324	4356	18225
Ratio Steps/Size	**r = s/z**	**64.06**	**10.25**	**3.16**	**0.24**	**0.06**

Figure 5.30: Ratio number of SA steps / size of Markov matrix. When r >> 1; there are plenty of steps to sample all the possible transitions, including those involving lowly populated conformations. When r ≈ 1; sampling is not so good, especially regarding lowly populated conformations, however, a reasonable full Markov matrix can still be obtained. When r < 1; sampling is definitely incomplete and only a Markov sub-matrix involving the highly populated conformations is trustworthy.

An alternative approach to handle this problem in a *reasonable* way to avoid the lack of information would be re-writing the Transition matrix gathering together all the lowly populated conformations under the label *conformational noise* and assigning it the normalised total probability of the individual conformations within the group. The result is a new *condensed* Markov matrix in the type of [Fig. 5.31]:

conf.	1	2	3	4	5	6	7	8	9	10	noise
1	0.1467	0.1133	0.1067	0.0867	0.1000	0.0600	0.0200	0.0133	0.0133	0.0267	0.3133
2	0.1458	0.1181	0.1042	0.1528	0.0764	0.0486	0.0278	0.0208	0.0208	0.0208	0.2639
3	0.2069	0.1724	0.1293	0.1121	0.0776	0.0517	0.0172	0.0086	0.0172	0.0000	0.2069
4	0.1373	0.1863	0.1275	0.0980	0.0686	0.0784	0.0098	0.0196	0.0196	0.0098	0.2451
5	0.2093	0.1279	0.0581	0.0233	0.0581	0.0233	0.0116	0.0233	0.0116	0.0116	0.4419
6	0.2041	0.2041	0.1429	0.0816	0.0816	0.0612	0.0408	0.0000	0.0000	0.0000	0.1837
7	0.1176	0.1176	0.0588	0.0000	0.1765	0.1176	0.0588	0.0000	0.0000	0.0000	0.3529
8	0.0625	0.1875	0.1250	0.2500	0.0625	0.0000	0.0000	0.0000	0.0000	0.0000	0.3125
9	0.1875	0.1250	0.0625	0.2500	0.1250	0.0000	0.0000	0.0000	0.0625	0.0000	0.1875
10	0.0667	0.1333	0.0000	0.1333	0.0000	0.0000	0.0000	0.1333	0.0000	0.0667	0.4667
noise	0.1353	0.1227	0.1377	0.0917	0.0654	0.0435	0.0055	0.0130	0.0152	0.0220	0.3480

Figure 5.31: Condensed Markov matrix for CD8 including main conformations and conformational noise. The last row is the summation of all the individual probabilities from conformation 11 to 135, and then divided by 125 –the number of conformations in the noise-. The last column is obtained, for any given conformation, subtracting to 1.0 the summation of all the transition probabilities from that very conformation to any of the first 10.

This condensed transition matrix is highly advisable and consistent with the flexible behaviour, also common to CD7.

5.2.1.5.4 Results

The system has –with high probability- no more than 6 important conformations, remaining most of the time in one of the first three or four ones; **1**, [**2Z,2S,2Z,2S**]; **2**, [**2Z,S,2Z,S,Z,S**]; **3**, [**2Z,2S,Z,S,Z,S**]; and **4**, [**Z,S,Z,S,Z,S,Z,S**].

The condensed transition matrix –the one including the conformational noise as a conformation- clearly describes the dual behaviour of CD8; that of the *main conformations* and that of the *conformational noise*. Under detailed examination, the 11x11 Markov submatrix [Fig. 5.31] shows that, similarly to CD7, there are several boxes in the lower triangle filled with probabilities other than zero, meaning that:

- Highly populated conformations not only interconvert into highly populated ones, but also into others less populated.
- Lowly populated conformations not only interconvert into highly populated conformations, but also into others less populated.

The last point is interesting since lowly populated conformations are now considered as a group. Therefore, the condensed Markov matrix points that:

- The transition probability from any of the *main conformations* towards one in the *conformational noise* ranges from 18% to 46%, which means about 1 in 3.
- The transition probability from any of the *main conformations* towards any of the main conformations ranges from 54% to 82%, which means nearly 2 in 3.
- The transition probability from *conformational noise* to any of the *main conformations* is nearly 65%, which means nearly 2 in 3.
- The transition probability of conformational noise to conformational noise is about 35%, which means nearly 1 in 3.

This result proves that, generally speaking, nearly 1 in 3 of the transitions involve conformations in the noise group; the *not-so-important* conformations weight about 30%, which actually means that –conformationally speaking- they are no longer second rate conformations but they are growing in importance.

The analysis confirms, on the one hand, the first six conformations as the principal ones, and on the other hand, the conformational noise group weighting about 33%.

5.2.1.6 Large Cyclodextrins: CD14, CD21, CD26 and CD28

The number of steps –1024- employed in the SA calculation proved to be inadequate for the large cyclodextrins as can be seen in [Fig. 5.32]. The slope is 1.0000 in all cases –CD21, CD26, CD28- except in CD14, where the tendency slightly changed to 0.9923; meaning that the saturation was about to begin.

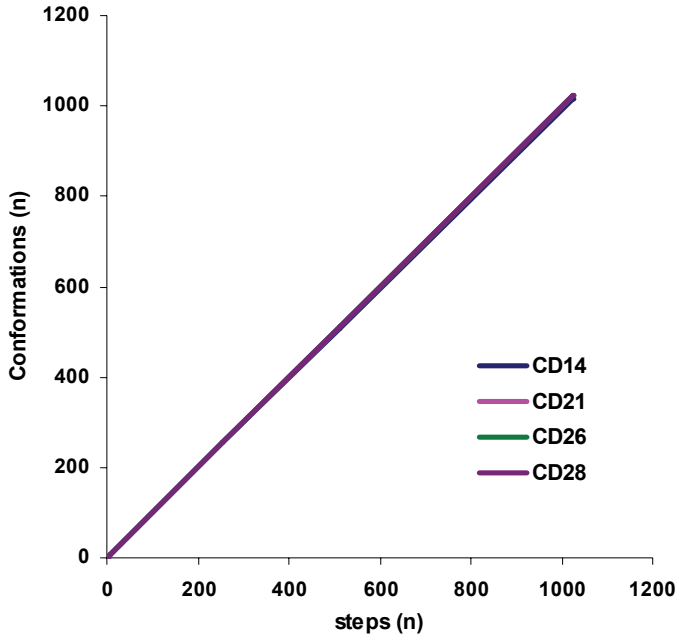


Figure 5.32: Plot of Number of Conformations vs. Number of SA steps for the Large Cyclodextrins. Conditions: Slow SA, 1024 steps, quadratic saturation analysis approach.

Anyway, almost in all cases the total number of conformations found was similar to the number of steps, which means that the number of steps was insufficient in comparison with the size of the Conformational Space of the Large CDs. Sampling was therefore ineffectively done and hence, it was useless to proceed with further analysis –Saturation Analysis, PCA and Markov- because the ensemble was not in thermodynamic equilibrium and not much information of interest could be derived:

- There were as many conformations found as SA steps.
- PCA showed that the conformational noise –a continuum of conformations in the conformational space- was the general behaviour detected in this group of

CDs; meaning that the extremely high flexibility makes the system to fold into a wide variety of conformations.

- Markov Analysis is useless since both the ratio, number of steps / Markov size, and the statistical weights, prove that the system is far from reaching thermodynamic equilibrium.

At this point, it was decided to carry out a new series of SA calculations for the Large Cyclodextrins –CD14, CD21, CD26 and CD28- in order to get an improved ensemble in the hope that thermodynamic equilibrium would be reached.

5.2.2 Fast SA [30 ps] 4096 steps + MM.

The series of 4096-steps SA calculations were carried out following the schema in [Fig. 5.33]. Every step included *"in vacuo"* conditions, 30 picoseconds, and external constraints to avoid chair-to-boat conformational interchanges in glucoses. Detailed information regarding *fast* SA conditions and further computational aspects can be found in the DVD: annex 10.1.1.2 (chapter X, AMBER 7 files).

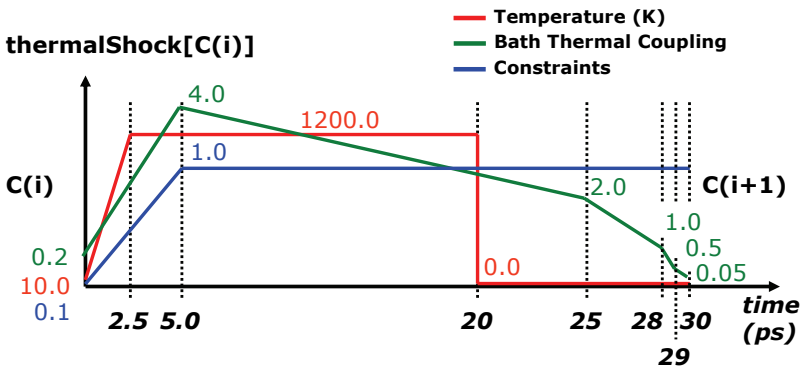


Figure 5.33: Schematic depiction of SA single 30-picoseconds step. Coloured lines, although not scaled- represent the dynamically coupled control parameters.

The new series of calculations involved a high number of steps, meaning longer simulation times. This new schema was computationally unaffordable maintaining the time length of the *slow* SA step, therefore, a new schema was proposed:

- An increment in the number of steps.
- A reduction of 90% of time in the SA single step: *fast* SA step.

The reduction of time from 300 ps –slow SA- to 30 ps –fast SA- in the SA step would render 10 times more conformations than the *slow* SA step running in the same amount of time. On the one hand, this modification would save computational time –nearly 90%- and, on the other hand, it would still be able to promote conformational changes. Nevertheless, a new problem was expected.

Both the *slow* SA and the *fast* SA steps heat the molecule to a peak of 1200 K; then, suddenly, it is cooled to 0 K and kept this way for a certain time. During this time, the molecule, depending on the thermal inertia of the system –bath thermal coupling, molecular size, and relaxing time at 0 K- descends into the closer energetic minimum. The question is, now that the time of the SA step –and the relaxing time at 0 K- has been reduced to 10% of the original time, will the molecule fully reach the energetic minimum? The answer, as can be seen in [Fig. 5.34], is; *no*.

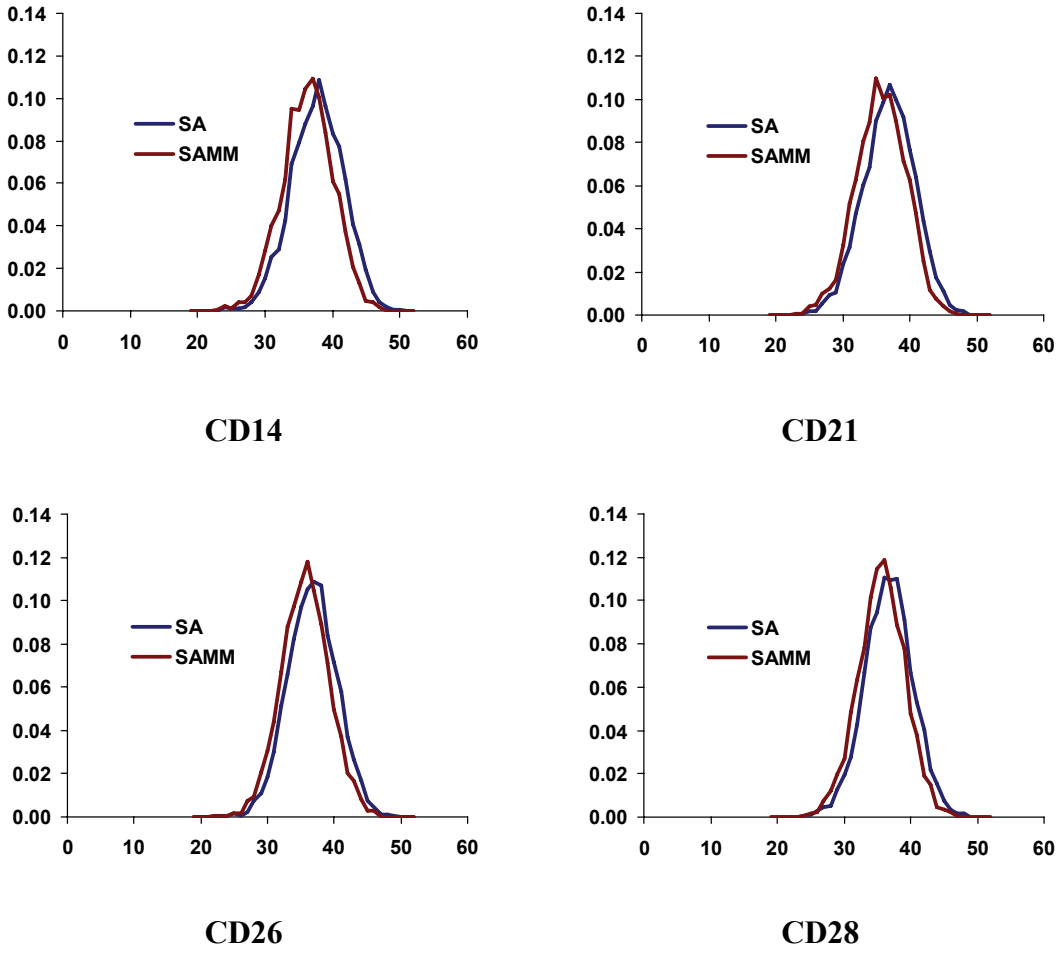


Figure 5.34: Normalised energy distributions for the Large Cyclodextrins. Frequency vs. Total Energy / Number of glucoses. Class width, ΔE = 1 kcal/mol. Distributions for the fast SA ensemble are represented in blue. Distributions for the SA ensemble subjected to 100 steps minimization, SAMM, are represented in brown.

For all the CDs, SAMM energy distributions are shifted to lower energies in comparison to SA distributions, meaning that conformations obtained by means of the *fast* SA process do not reach the energetic minimum and, therefore, a subsequent MM treatment after SA is required to ensure full convergence.

5.2.2.1 Large Cyclodextrins: CD14, CD21, CD26 and CD28

Unfortunately, the situation did not improve by extending to 4096 the number of steps employed in the SA calculation. Similarly to the case of 1024 steps, the new schema of 4096 SA steps followed by MM, although better than the original SA, still proved to be inadequate for the large cyclodextrins, as can be seen in [Fig. 5.35]. The slope is again 1.0000 in all cases –CD21, CD26, CD28- except in CD14, where the tendency changed to 0.9462; which is less than 0.9923 and means that the saturation process was better than in the case of 1024 steps, but still insufficient.

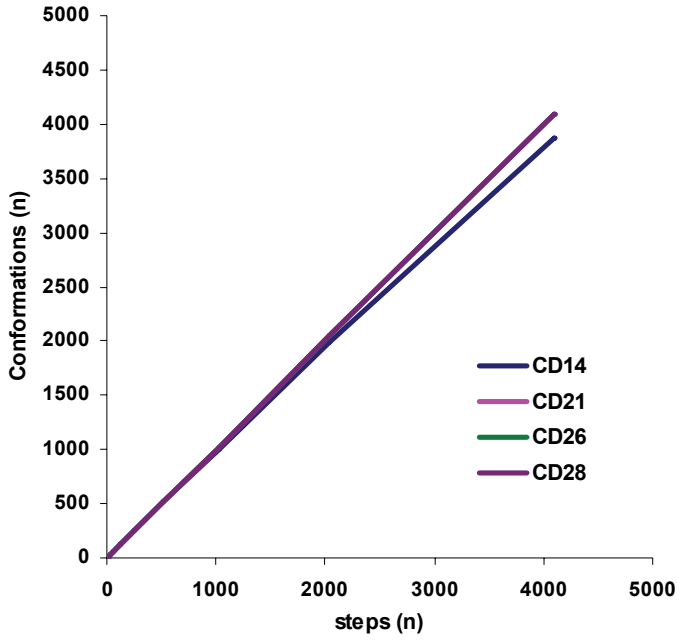


Figure 5.35: Plot of Number of Conformations vs. Number of SA steps for the Large Cyclodextrins. Conditions: Fast SA, 4096 steps, quadratic saturation analysis approach.

Similarly to the 1024 steps SA process, in almost all cases the total number of conformations found was similar to the number of steps, meaning that the number of steps was insufficient in comparison to the size of the Conformational Space of the Large CDs. Sampling was ineffectively done and the ensemble was not in thermodynamic equilibrium. Hence, it was useless to proceed with further analysis –

Saturation Analysis, PCA and Markov- and not much information of interest could be derived:

- There were as many conformations found as SA steps.
- PCA showed that the conformational noise was the general behaviour detected in Large CDs, the system folding into a wide variety of conformations because of the extremely high flexibility.
- Markov Analysis is useless since all the ratio, number of steps / Markov size, and the statistical weights, prove that the system is far from reaching thermodynamic equilibrium.

The set of results obtained for large CDs –from both 1024 and 4096 SA steps- suggests that:

- An *extremely large* amount of steps is required for a full sampling.
- Further SA calculations should be avoided until new information about the size of the Conformational Space could be obtained by other means.

The only macromolecule in the ensemble that remotely seems to approach saturation and thermodynamic equilibrium is CD14, while CD21, CD26 and CD28 are absolutely undersampled.

5.3 GROUP PROPERTIES

The point of testing our conformational search methodology with series of cyclodextrins is interesting not only for the methodological results but also for the profitable information about cyclodextrin ensembles in thermodynamic equilibrium; valuable results can be obtained from the SA groups of conformations. In this sense, it is important not to forget that:

- Small cyclodextrins –CD5, CD5x, CD6, CD7, and CD8- were effectively sampled, meaning that all the information regarding these cyclodextrins is contained in those sets of molecules.
- Large cyclodextrins –CD14, CD21, CD26 and CD28- were presumably ineffectively sampled. However, the other possibility is that the behaviour of

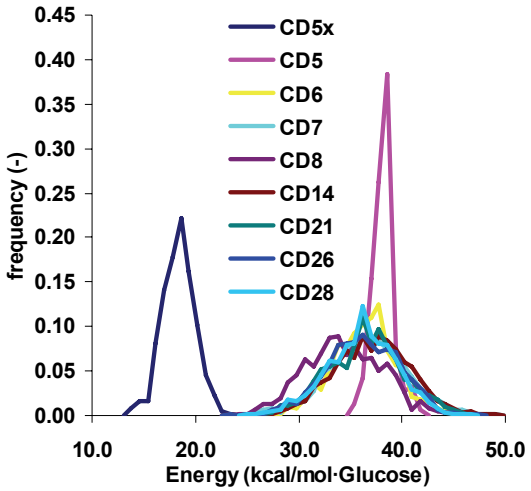
those CDs is *different* because they are such a flexible group of macromolecules that it is impossible to perform a sampling employing the same *frame* used for the small ones. What if the result of a conformational search under 10^{10} steps would have been the same? This would probably mean that 5000 or 10^{10} steps would render similar information and therefore we could still study some properties of those macromolecules –not those involving conformational information- on the basis of the 4097 SA ensembles.

Therefore, the whole set of cyclodextrins in 1024 and 4097 ensembles will be compared in the next sections. Since molecular mechanics do not allow direct comparisons involving systems containing a different number and type of atoms –particularly when comparing energies-, results included in the next sections –unless special mention- are presented already normalised to the total number of glucoses. The analysis of some selected parameters: energy, molecular surface area, radius of gyration, hydrogen bonding and other representations like molecular superimpositions and average structures is presented.

5.3.1 Histograms of Energy

The analysis of the Energy distributions for the 1024 SA ensembles is shown in [Fig. 5.36]. All cyclodextrins were studied under the same conditions so they can be compared. Two important properties are depicted:

- All distributions are centred in the same energetic values; however, the distribution widths grow as the number of glucoses is bigger. This result is a direct consequence of the flexibility and also an alternative proof to the fact that flexibility grows in parallel with the number of glucoses.
- The average energetic value of CD5x is dramatically different from the value of the other cyclodextrins; CD5, CD6, CD7, CD8, CD14, CD21, CD26, and CD28. The “x” value is nearly half the values of the “i”, so it cannot be regarded just as a fluctuation in the ensemble.



1025SA	glucoses	Energy/Glu (kcal/mol.Glu) average	std.dev.
CD5x	5	17.98	1.54
CD5	5	37.79	1.04
CD6	6	35.99	3.03
CD7	7	36.24	3.90
CD8	8	33.99	3.78
CD14	14	36.70	3.70
CD21	21	36.05	3.61
CD26	26	35.78	3.55
CD28	28	35.69	3.37

(a) (b)

Figure 5.36: Full series of cyclodextrins. 1025 SA Ensembles. (a) Normalised total energy distributions. Class width, ΔE = 0.8 kcal/mol·Glucose. (b) Table of average energies and standard deviations.

After detailed examination of the energy decomposition, it was detected that the strange behaviour was originated by the Electrostatic terms [Fig. 5.37]. All the energetic rates having in common the set of charges by Dr. Itziar Maestre fall in a similar tendency, overlapping almost completely; nevertheless, the rates involving the species with the set of charges by Dr. Javier Pérez have an acutely different behaviour in the keys 6 and 8.

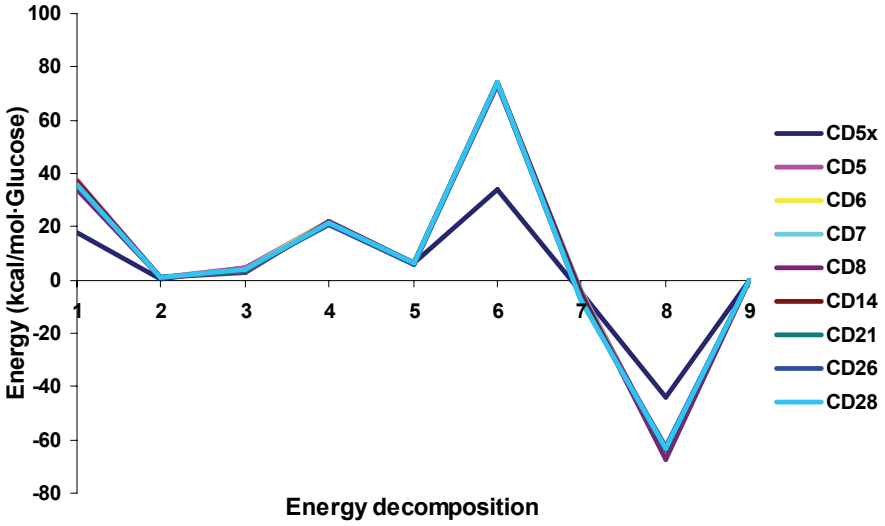
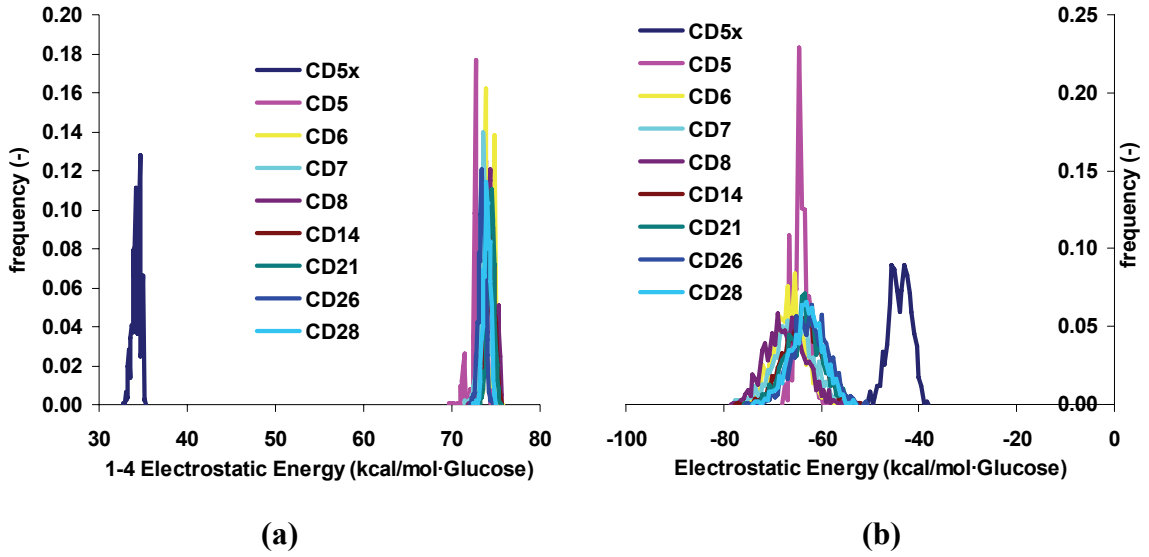


Figure 5.37: Total Energy and decomposition in terms: 1 = Total Energy; 2 = Bond Energy; 3 = Angle Energy; 4 = Dihedral Energy; 5 = 1-4 Non-bonding Interactions; 6 = 1-4 Electrostatic Interactions; 7 = Van der Waals Energy; 8 = Electrostatic Energy; 9 = Restraint Energy.

All the distributions of the individual energetic terms of the decomposition were plotted and studied. Then, the 1-4 electrostatic interactions and the electrostatic energy were again confirmed to be the responsible ones for the special energetic values [Fig. 5.38].



(a) **(b)**

Figure 5.38: Full series of cyclodextrins. 1025 SA Ensembles. (a) Normalised 1-4 electrostatic energy distributions. Class width, ΔE = 0.1 kcal/mol·Glucose. (b) Normalised electrostatic energy distributions. Class width, ΔE = 0.5 kcal/mol·Glucose.

The differences in energies are significant. The 1-4 electrostatic energy is half the value of the "i" cyclodextrins; a gap of nearly 35 kcal/mol·Glucose. In the case of the electrostatic energy the difference is 20 kcal/mol·Glucose. Both 1-4 electrostatic energy and electrostatic energy weigh considerably in the total energy, so important variations in their values mean important changes in the total energy.

It has been proven that using different sets of charges in the units deeply affects the results in the calculations:

- Conformational behaviour is modified, as explained when the results of CD5x were shown and compared with those of CD5.
- Energetic values are also modified, as the energetic distributions state beyond any doubt.

The importance of using an appropriate, consistent and trustworthy set of charges was already noticed and stressed by Dr. Miguel de Federico[189] and later by Dr. Javier Pérez[190]. Both of them faced similar situations working with AMBER 7 and studied this problem in depth. In this sense, the results presented in this work come to support their Theses.

Energy comparisons involving 1025SA, 4097SA and 4097SAMM ensembles are omitted since they have been already explained.

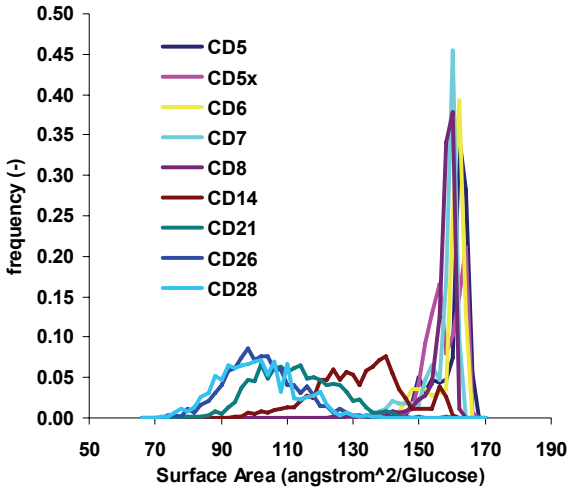
5.3.2 Surface Area

Molecular surface is indirectly obtained in AMBER 7 by means of MM-PBSA module (check AMBER 7 manual). It is one of the parameters necessary for evaluating the solvation energy by MM-PBSA methodology and the solvent accessible surface –the parameter considered here as the molecular surface- is dumped in a production file.

Molecular surface graphics help understanding flexibility and folding. On the one hand, it is not difficult to realise from [Fig. 5.39] that cyclodextrins show two different behaviours regarding flexibility:

- Small CDs –CD5, CD6, CD7, and CD8- have sharp and narrow distributions, which is in accordance with a small number of rigid conformations.
- Large CDs –CD14, CD21, CD26, and CD28- have scattered and wide distributions, pointing that there are many flexible conformations.

[189] De Federico, M.A.; *Doctoral Thesis*. Chemistry Department. UAB. **2006**.
[190] Pérez, J.; *Doctoral Thesis*. Chemistry Department. UAB. **2008**.

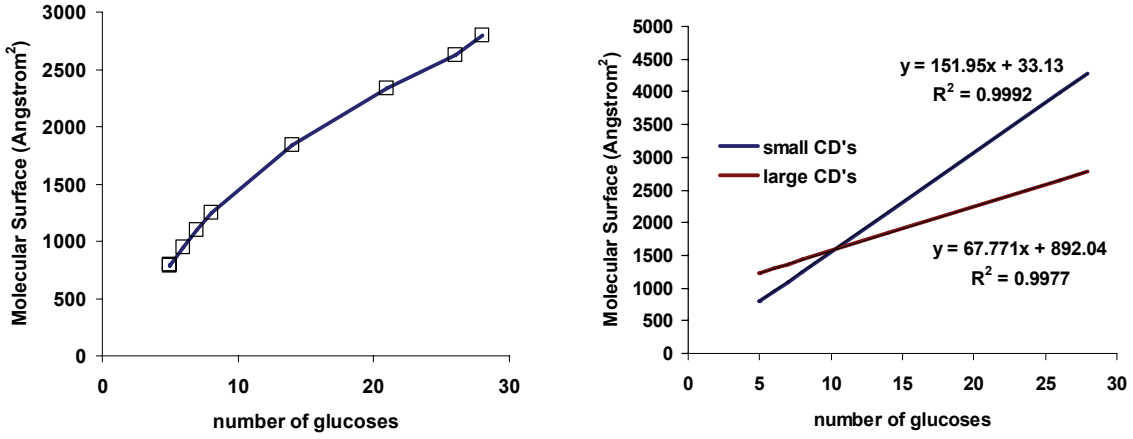


1025SA	glucoses	Mol.Surf./Glu (Ang^2/Glu) average	std.dev.
CD5	5	159.54	4.48
CD5x	5	157.16	4.66
CD6	6	158.39	4.64
CD7	7	156.05	5.37
CD8	8	156.23	4.27
CD14	14	130.99	13.16
CD21	21	111.26	12.03
CD26	26	101.11	10.52
CD28	28	100.02	11.48

(a) **(b)**

Figure 5.39: Full series of cyclodextrins. 1025 SA Ensembles. (a) Normalised surface area distributions. Class width, ΔA = 2.0 Angstrom^2/Glucose. (b) Table of average surface areas and standard deviations.

On the other hand, there is the contraction phenomenon in the normalised surface area. As the number of glucoses grows, the surface area per unit of glucose diminishes, meaning that there is not a linear correspondence between number of glucoses and molecular surface [Fig. 5.39b and 5.40a].



(a) **(b)**

Figure 5.40: Full series of cyclodextrins. 1025 SA Ensembles. (a) Graphic of the molecular surface area (not normalised to the number of glucoses) vs. number of glucoses. (b) Linear tendencies; the intersection delimits the frontier between CDs that cannot fold –in blue- and those that can fold –in red-.

In fact, this relationship is more complex: the small CDs seem to follow a different linear tendency than the large ones. Linear regressions [Fig. 5.40b] confirmed this assumption and the graphical plot helped also to identify the frontier between the rigid

CDs that cannot fold and the flexible ones –exhibiting secondary structure- that can fold and therefore behave as foldamers. This frontier –the intersection point- is located in cyclodextrins with 10 glucoses, which completely agrees with the experimental results by Saenger and coworkers[191] already mentioned at the end of Chapter III.

5.3.3 Radius of Gyration

The radius of gyration is a well-known molecular descriptor that estimates the size of a molecule of any shape. The IUPAC –in *The Golden Book*- officially gives the next definition[192]: for a rigid particle consisting of mass elements of mass, $\mathbf{m_i}$, each located at a distance, $\mathbf{r_i}$, from the centre of mass, the radius of gyration, **s**, is defined as the square root of the mass-average of for all the mass elements [Eq. 5.2]:

$$s = \sqrt{\frac{\sum_i m_i r_i^2}{\sum_i m_i}}$$

Equation 5.2: Definition of the Radius of Gyration by IUPAC.

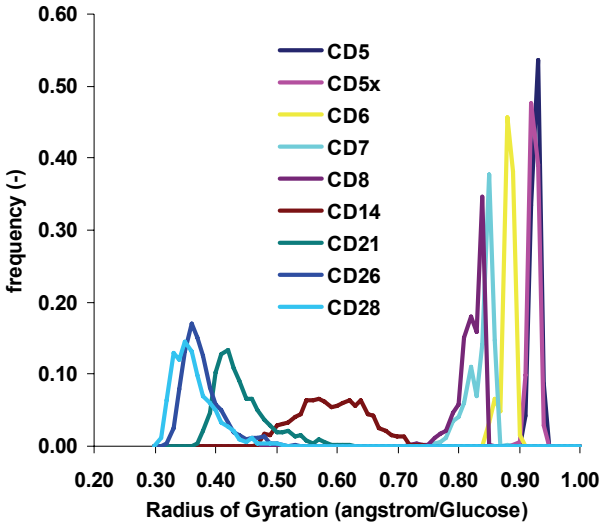
The radius of gyration is obtained in AMBER 7 by means of CARNAL module (check AMBER 7 manual) and the command file can be consulted in the DVD: appendix 10.2.2.2 (chapter X Amber 7 files).

Results derived from the radius of gyration come to confirm, from an alternative point of view, the conclusions about flexibility and folding already stated in the last section. As can be seen, distributions are better defined [Fig. 5.41] than in the case of molecular surface, which is good for our purposes. Focusing on flexibility, we find the dual behaviour of the cyclodextrins depending on the group they fall in:

- Small CDs –CD5, CD6, CD7, and CD8-. Their distributions are sharp and narrow –in accordance with a small number of rigid conformations-.

[191] Saenger, W.; Jacob, J.; Gessler, K.; Steiner, T.; Hoffmann, D.; Sanbe, H.; Koizumi, K.; Smith, S.M.; Takaha, T.; *Chem. Rev*. **1998**. *98(5)*. 1787-1802.
[192] (a) Gold, V.; Loening, K.L.; McNaught, A.D. and Shemi, P. *The Gold Book: Compendium of Chemical Terminology*. Blackwell Science, **1987**. ISBN 0-63201-7651(8). (b) McNaught, A.D. and Wilkinson, A. *The Gold Book: Compendium of Chemical Terminology, 2nd edition*. Blackwell Science, **1997**. ISBN 0-86542-6848. (c) http://old.iupac.org/publications/compendium/index.html. *"The Gold Book"* on-line.

- Large CDs –CD14, CD21, CD26, and CD28. Their distributions are scattered and wide –compatible with many flexible conformations-.

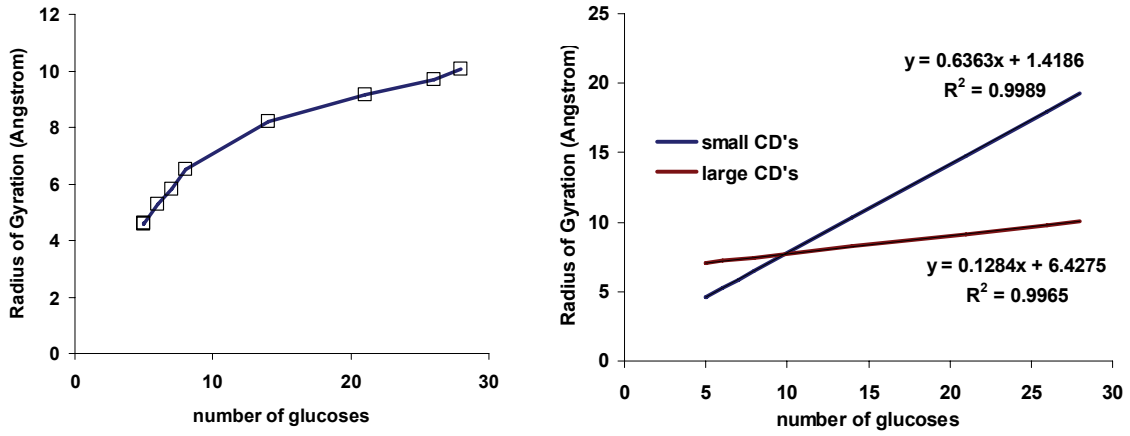


1025SA	glucoses	Rad.Gyr./Glu (Ang/Glu) average	std.dev.
CD5	5	0.9216	0.0065
CD5x	5	0.9183	0.0067
CD6	6	0.8766	0.0097
CD7	7	0.8326	0.0206
CD8	8	0.8163	0.0198
CD14	14	0.5866	0.0552
CD21	21	0.4365	0.0426
CD26	26	0.3732	0.0334
CD28	28	0.3591	0.0342

(a) **(b)**

Figure 5.41: Full series of cyclodextrins. 1025 SA Ensembles. (a) Normalised radius of gyration distributions. Class width, ΔRadGyr = 0.01 Angstrom/Glucose. (b) Table of average radius of gyration and standard deviations.

Regarding folding, it is interesting to notice that all the distributions for the normalised radius of gyration are strictly shifted towards lower values as the number of glucoses grows [Fig. 5.41b]; that is the contraction phenomenon previously mentioned, meaning that there is no linear correspondence between the number of glucoses and the radius of gyration [Fig. 5.42a].



(a) **(b)**

Figure 5.42: Full series of cyclodextrins. 1025 SA Ensembles. (a) Graphic of the radius of gyration (not normalised to the number of glucoses) vs. number of glucoses. (b) Linear tendencies; the intersection delimits the frontier between CDs that cannot fold –in blue- and those that can fold –in red-.

The same situation explained for the surface area is faced here; the small CDs follow a different linear tendency than the large ones. Linear regressions [Fig. 5.42b] were calculated for each group and the graphical plot helped to fix the frontier between the rigid CDs that cannot fold and the flexible CDs exhibiting secondary structure that can fold and therefore behave as foldamers.

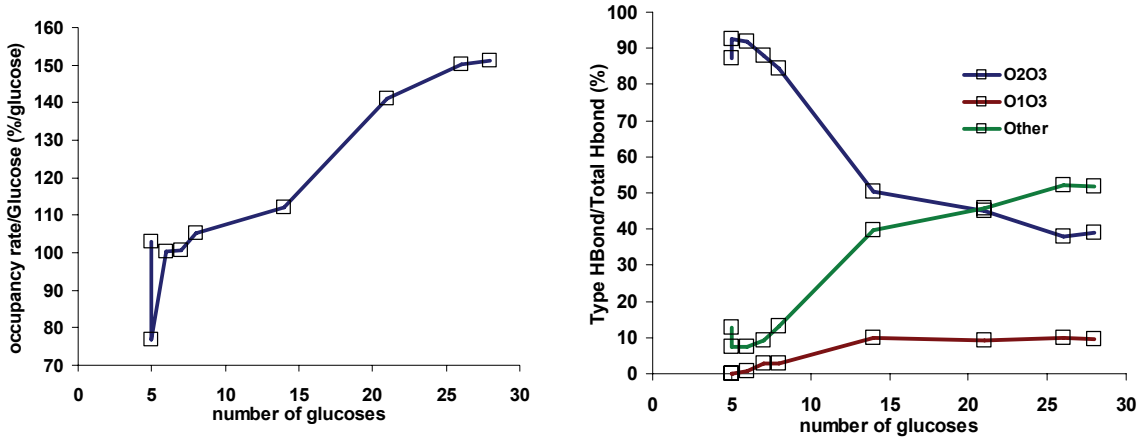
The intersection point was located also in cyclodextrins with 10 glucoses; the same result was obtained from the molecular surface intersection, and is also in complete agreement with the already mentioned experimental results by Saenger.

5.3.4 Hydrogen Bonding

The graphics including hydrogen bonding are useful since they supply information about intramolecular interactions, and flexibility.
The Hydrogen Bonding is obtained in AMBER 7 by means of CARNAL and ptraj modules (check AMBER 7 manual) and both the command files can be consulted in the DVD: appendices 10.2.1.2 –ptraj- and 10.2.2.1 –CARNAL- (chapter X Amber 7 files). The whole group of conformations in the 1025 SA and 4097 SA/SAMM chains were considered for this analysis, however, only results from 1025 SA ensembles are included since 4097 ensembles behave similarly.

It is not difficult to realise, from [Fig. 5.43a], that the intramolecular interactions are favoured when flexibility grows. The small cyclodextrins present a similar behaviour with the exception of CD5x, which falls below the group. This result is understandable because the set of charges –as has been already shown- has a direct influence in the conformational freedom, allowing different stable conformations to be found apart from those mainly stabilised by hydrogen bonding interactions. The overall tendency in this group is that they have comparable occupancy ratios, however CD6, and CD7 seem to be under CD5 and CD8, which is surprising since a linear tendency was expected –further studies would be of interest to give an appropriate answer-.

The situation in the case of large cyclodextrins is different; the intramolecular interactions increase as the total number of glucoses grows. The question is; how does this phenomenon happen? The answer can be found in [Fig. 5.43b].

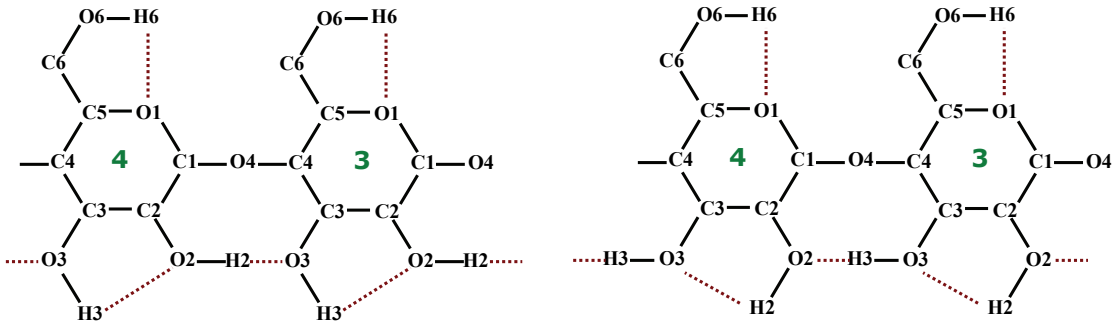


(a) (b)

Figure 5.43: Full series of cyclodextrins. 1025 SA Ensembles. Hydrogen bond statistics: (a) Normalised occupancy rate of the total hydrogen bonds (the scale ranges from 0 to 300). (b) Percentage of hydrogen bonds involving secondary hydroxyls –in blue-, inverted glucoses –in red- and other types –in green-.

The graphic includes the decomposition of the total hydrogen bonds into two of the main types, and the other ones:

- Interactions involving hydroxyl groups O2-H2 and O3-H3 –represented O2O3-.
- Interactions involving hydroxyl groups O3H3 and O1 –represented O1O3-.
- Other hydrogen bonding interactions.



(a) (b)

Figure 5.44: Schema of the two types of hydrogen bonds involving secondary hydroxyls; the most important ones in cyclodextrins. (a) Hydroxyls pointing to the n-1 direction [O2-H2···O3-H3]. (b) Hydroxyls pointing to the n+1 direction [O3-H3···O2-H2].

In rigid cyclodextrins, the hydrogen bonds involving secondary hydroxyls –O2O3- [Fig. 5.44] are the most important ones and weight more than 85% of the total. This circular uninterrupted chain of intramolecular hydrogen bonds stabilise the molecules and prove that glucose inversions within the macrorings are, at most, highly infrequent. Other interactions involving chains of primary hydroxyls are detected; however, they are poorly populated since the distance between atoms is bigger. Nevertheless, interactions involving O6H6 hydroxyl with O1 oxygen are quite frequent, as can be seen in the average structures in section 5.3.5.2.

The behaviour of large cyclodextrins is different [Fig. 5.43b]. There is a decrease in the O2O3 hydrogen bonding –meaning that the chain of secondary hydroxyls is broken more frequently- but it is balanced by the increase in the general hydrogen bonds –in green-. This new type of hydrogen bonding interactions detected between non-adjacent glucose units proves that the only way to achieve these interactions is by means of folding; which makes possible that glucoses separated in the chain –primary structure- meet in the space. Furthermore, the increase in the O1O3 ratio proves that glucose inversion [Fig. 5.45] is quite a common phenomenon within this group of macrorings, rising up to 10% of the total glucoses.

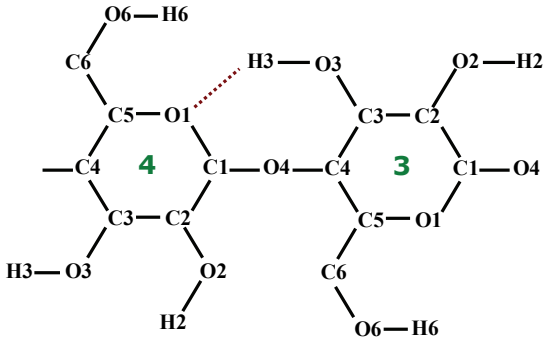


Figure 5.45: Schema of hydrogen bond of the type [O3-H3···O1]. This interaction is only possible when inverted glucoses are present in the cyclodextrin.

Hydrogen bonding analysis has proved to be a powerful tool and, though it has been roughly used in this section, going beyond this point would be of interest if further *in-detail* studies were done.

5.3.5 The rigid & flexible families

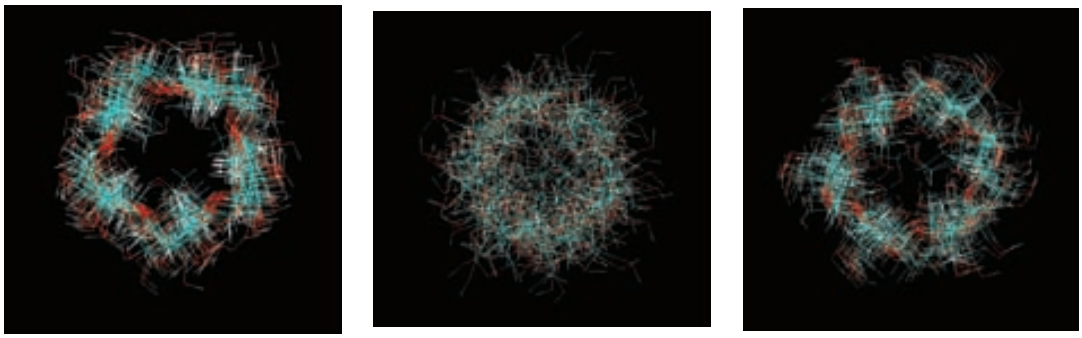
Until this point, it has been said that cyclodextrins more or less fall into two categories; the rigid ones and the flexible ones. In this sense, some analytical proof has been already presented: several data from conformational search –saturation analysis, PCA and Markov analysis-, distributions –energetic, solvent accessible surface and radius of gyration- and information derived from hydrogen bonds. However, none of them were visual or intuitive enough despite of their mathematical consistency.

This section approaches the flexibility from a graphical viewpoint, taking profit of two graphical representations: the structure superposition and the average structures.

5.3.5.1 Structure Superposition

The structure superposition of a set of conformations is a plot representing a cluster of conformations. When certain conformations weight more than others, their contribution to the cluster is noticeable and, therefore, the image gives a hint about the flexibility and the fluctuations of the ensemble.

The images in [Fig. 5.46] were created with ***VMD molecular viewer***[193]. The individual conformations in the SA ensembles were glued together with AMBER 7 module ptraj into a single trajectory file containing the Markov chain. The structures were centred –on the centre of mass- and oriented minimising RMSD –root medium square distance- to the previous conformation. The command file can be consulted in the DVD: appendix 10.3.1.3 (chapter X, Amber 7 files).



CD5 **CD5x** **CD6**

[193] (a) http://www.ks.uiuc.edu/Research/vmd/ (b) Humphrey, W.; Dalke, A.; Schulten, K.; *J. Mol. Graphics*. **1996**. *14(1)*. 33-38.

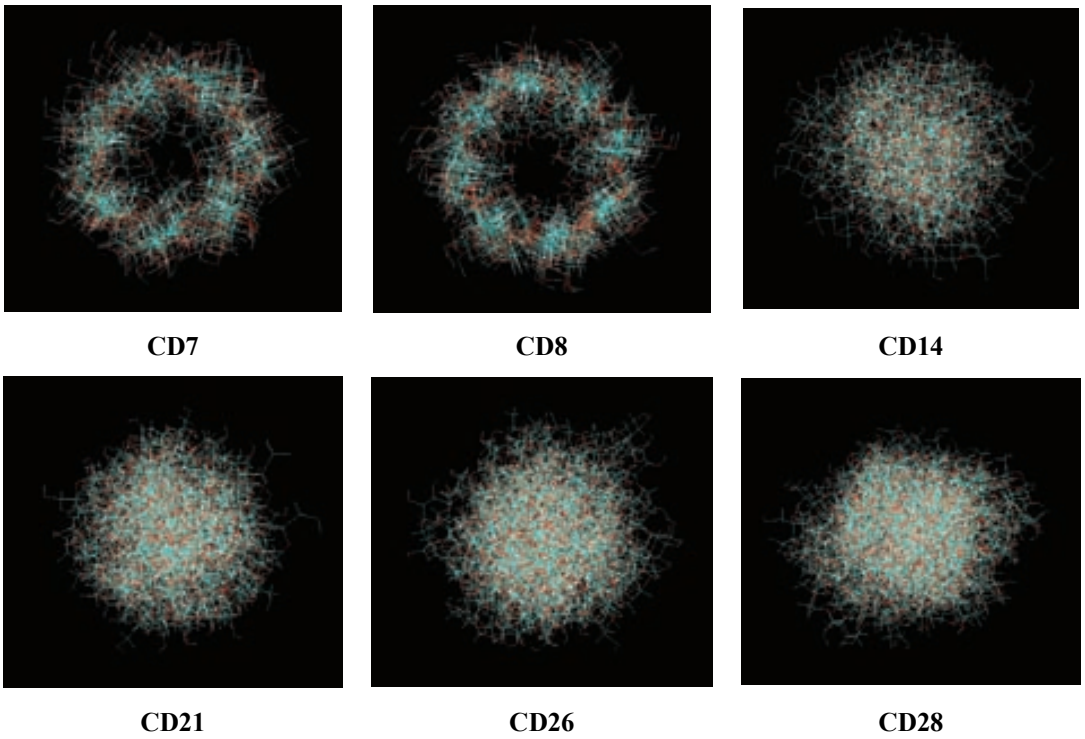


Figure 5.46: Superimposed structures of cyclodextrins. CD5, CD5x, CD6, CD7 and CD8 images obtained from 1025SA ensemble with frequency = 1/15. CD14, CD21, CD26 and CD28 obtained from 4097SAMM ensemble with frequency = 1/200.

The pictures clearly show the conformational behaviour of the cyclodextrins:

- Small CDs: CD5, CD5x, CD6, CD7 and CD8. The well defined clusters depicted prove that they are mainly rigid species and there are some conformations that weight predominantly in the group. This is also connected with the fact that the cavity is maintained in all cases –it does not collapse-. It is also observed that the flexibility increases as the number of glucoses grows; which can be related with the rising fuzziness of the images. In this sense, it is important to notice that there is a significant difference in flexibility between CD5 and CD5x, which is in agreement with the conformational results for both cyclodextrins. Otherwise, the "i" series continually grows in flexibility from rigid species –CD5 and CD6- to semi-rigid ones –CD7 and CD8-.

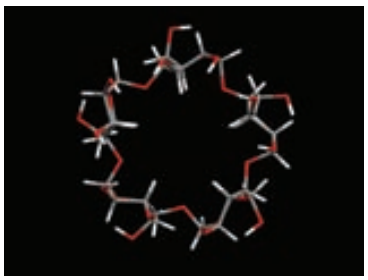
- Large CDs: CD14, CD21, CD26 and CD28. The situation is now dramatically different in comparison with the small CDs. The images show the result of superposing *conformational noise*. The high rate of fuzziness points out the extremely high flexibility in this group. Besides, no predominant conformations

or cavities were found, meaning that large numbers of conformations are almost equally stable close to the minimum within a reasonable energetic window.

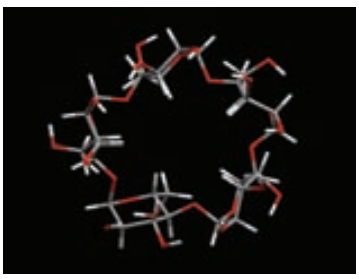
5.3.5.2 Average structures

The average structure of a set of conformations is an *artificial* structure which mostly represents their overall behaviour. When certain conformations weight more than others, their contribution to the average structure is noticeable and, in an important degree, the average structure takes a lot from them. Therefore, the average structure is probably the most similar conformation to the main conformation.

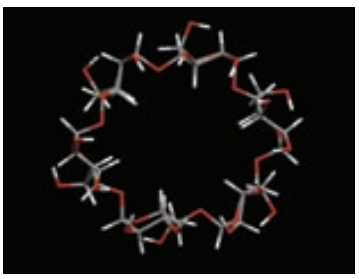
The images in [Fig. 5.47] were created with ***Mercury molecular viewer***[194]. A PDB file containing the average structure was previously computed with AMBER 7 module ptraj processing the trajectory file of SA Markov chain conformations already used for the structure superposition. The command file can be consulted in the DVD: appendix 10.3.1.2 (chapter X Amber 7 files).



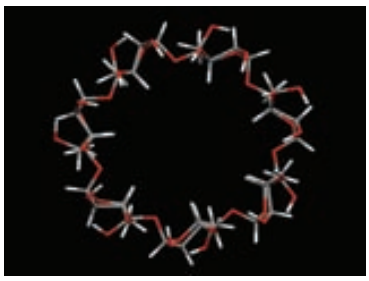
CD5



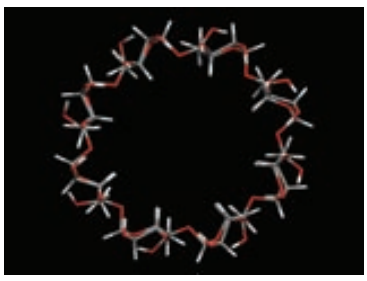
CD5x



CD6



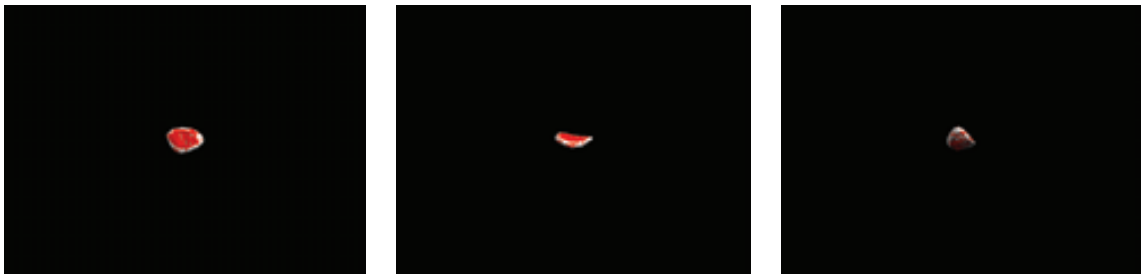
CD7



CD8

CD14

[194] (a) **Mercury - *Crystal Structure Visualisation and Exploration Made Easy***. Copyright © 2004-2009 The Cambridge Crystallographic Data Centre. 12 Union Road, Cambridge, CB2 1EZ, UK, +44 1223 336408. *Registered in England No.2155347. Registered Charity No.800579.*
(b) http://www.ccdc.cam.ac.uk/mercury/ (c) http://www.ccdc.cam.ac.uk/free_services/mercury/



CD21 **CD26** **CD28**

Figure 5.47: Average structures of cyclodextrins. CD5, CD5x, CD6, CD7 and CD8 images obtained from 1025SA ensemble. CD14, CD21, CD26 and CD28 obtained from 4097SAMM ensemble.

The pictures in [Fig. 5.47] are tightly correlated to that of [Fig. 5.46]. And again, an important difference is found between small and large cyclodextrins:

- The average structures of the small cyclodextrins are *plausible* conformations, meaning that they make sense by themselves: macroring toroidal shape, cavity strongly maintained, glucoses in an *all-Z/S* disposition, no inverted glucoses, close continuous chain of hydrogen bonds involving all the secondary hydroxyls, and primary hydroxyls pointing to the oxygen within the glucose.

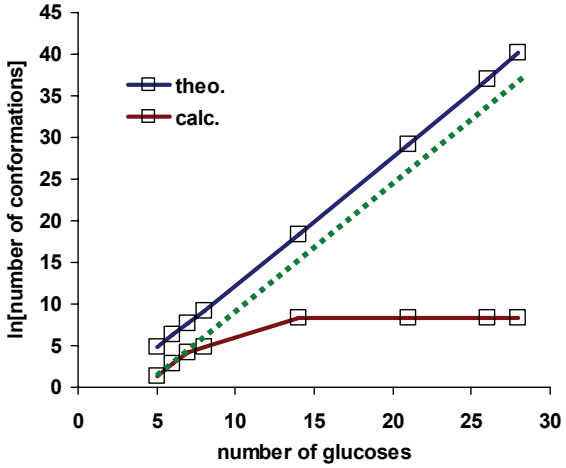
- Nevertheless, the average structures of the large cyclodextrins are *artefacts* that have nothing to do with the average conformational behaviour of the ensembles; however, the results still make sense in a certain way. The average structures obtained were certainly strange since all the atoms in the molecules were placed within a distance of nearly 1 angstrom from the centre of mass. How is it possible that such a group of large macromolecules could be reduced to a cumulus of dots in the vicinity of the origin of coordinates? The answer is that the enormous flexibility makes nearly all the conformations almost equal in probability and, therefore, there is not a set of main conformations that could influence in the average as was the case in the small CDs. The conformational space of all the large cyclodextrins is huge and only *conformational noise* was obtained when the conformational search was carried out. The images represent the average of a *conformational noise* in the same way that the average of a set of random data is a value that has nothing in common with the original data, and hence, does not represent the set of values. In fact, the sets of conformations are uncorrelated and cannot be represented by a single conformation.

5.3.6 Saturation Approach and Equation of Conformations

Chapter IV included a theoretical section where some equations for predicting the total number of conformations for cyclodextrins were developed. In this chapter, on the basis of the Saturation Analysis, the total number of conformations has been determined for the small cyclodextrins –at least for CD5, CD5x and CD6-. In this situation, the natural question rises; is there any relationship between the theoretical and SA ensembles –experimental- number of conformations?

This section is the answer to that question and also an attempt to estimate the necessary amount of steps for a reasonable sampling –if new SA process were done in the future- combining both the theoretical and computational results.

The results in [Fig. 5.48] are quite straightforward. On the one hand, it seems that theoretical –in blue- and computational –in red- results for small cyclodextrins correlate extremely well. As can be seen, the tendency slightly changes in the case of CD8, which in the end was the poorly sampled one within the group of the small cyclodextrins. The behaviour, hence, separates from the expected tendency –in green-.



(a)

glu.	Equation ln[C(h,n)]	SA/SAMM [C]	SA/SAMM ln[C]	Estimated Steps [C]	Estimated Steps ln[C]
5	4.83	4	1.39	4	1.39
6	6.26	18	2.89	19	2.93
7	7.71	66	4.19	87	4.46
8	9.19	135	4.91	404	6.00
14	18.28	3870	8.26	4.1E+06	15.22
21	29.14	4097	8.32	1.9E+11	25.98
26	36.98	4097	8.32	4.2E+14	33.67
28	40.12	4097	8.32	9.0E+15	36.74

(b)

Figure 5.48: (a) Graphic of the natural logarithm of the Total Number of Conformations vs. number of glucoses; theoretical from [Eq. 4.2] –in blue- and obtained from 1025SA and 4097SAMM calculations –in red-. Dotted line in green represents the expected tendency. (b) Table of number of conformations; last column is data for green dotted line.

Making predictions out of the range of calibration is always discouraged –only the small CDs were employed in the correlation- so the dotted green line should only be regarded as an aid in order to evaluate the number of steps for future calculations rather than a *solid* law.

As can be seen in [Fig. 5.48a], apparently the large cyclodextrins were badly sampled. In principle, 4096 steps seem to be insufficient for a reasonable conformational study; anyway, it does not seem to be the only conclusion. In fact, several possibilities are considered:

- The problem of sampling is just a matter of steps. Longer SA simulations will solve the problem.
- The conformational space is large enough for allowing multiple conformations on the basis of energetic criteria. Large cyclodextrins do not really have groups of *main conformations* like the small ones but they behave as a pack of *conformational noise*; hence, longer simulations will render similar results –just more *conformational noise*- and therefore 4097 conformations are enough for describing these systems.
- The Descriptor developed in this work, D-II, is not the appropriate one to describe by itself large cyclodextrins because it does not take into account *folding* and this lack of information could be important. New descriptors like *2D distance matrices* between glucoses or *polynomials* that describe the line of the macroring in space should complement information supplied by D-II.

5.4 SUMMARY OF RESULTS

This section contains the most important results described in this chapter:

Regarding Methodology:

- Saturation Analysis has proved to be a powerful and easy handling conformational search tool. In combination with PCA and Markov Analysis, it supplies valuable conformational information.

- Fast SA followed by MM protocol is more advisable than only Slow SA protocol. It ensures better convergence ratios to the minima and saves considerable amounts of time.

Regarding Cyclodextrins:

- The Conformational Space of the Cyclodextrins can be divided into two groups:
  - Rigid and Semirigid [Fig. 5.49a]: Well defined conformations.
    - ❖ CD5x, CD5 and CD6: *Main conformations*.
    - ❖ CD7 and CD8: *Main conformations* and *Conformational Noise*.
  - Flexible [Fig. 5.49b]: Not defined conformations.
    - ❖ CD10 and larger: *Conformational Noise*.

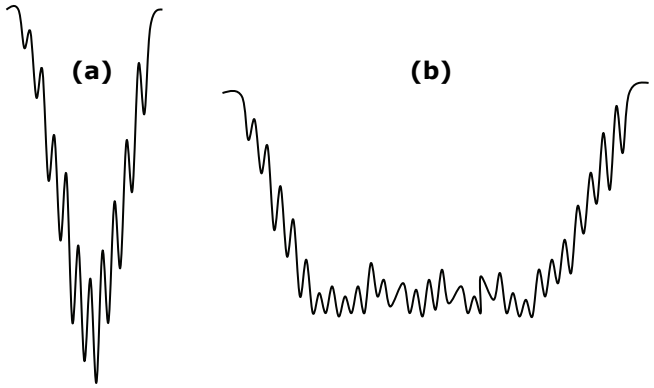


Figure 5.49: Schema of the conformational spaces for (a) small and (b) large cyclodextrins.

- The cyclodextrin containing 5 glucoses is the single one that cannot invert glucoses due to its extremely high rigidity.
- Flexibility grows as the number of glucoses grows.
- The employ of different sets of charges in the molecules strongly affects the conformational behaviour of the system.
- New descriptors like *2D distance matrices* or *polynomials* that describe the secondary structure should complement information supplied by D-II.

5.5 STRUCTURE SELECTION FOR MOLECULAR DYNAMICS

The conformational search not only rendered plenty of results that allowed a study in depth of the ensembles but also was the starting point of the second part of this work. Some conformations for every cyclodextrin were selected from the ensembles:

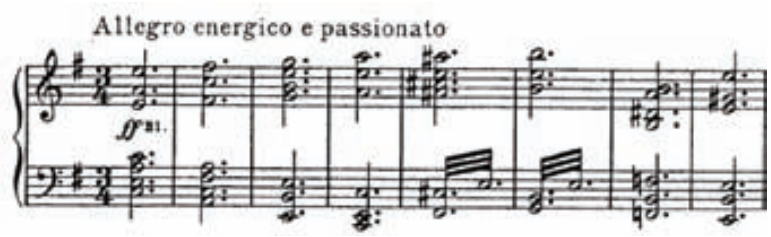
- 1025 SA Markov chain: CD5, CD5x, CD6, CD7 and CD8, and
- 4097 SAMM Markov chain: CD14, CD21, CD26 and CD28.

The idea was running several MD calculations to check the alleged dependency of the trajectories to the starting conformation. In this sense, by the time this series of calculations were planned, both D-I and D-II were still under evaluation and hence the selection criteria was undertaken on the grounds of D-I ratios [Fig. 5.50].

MD Conf.		SA/SAMM number	Selection Criterion	D-I weight (%)	D-I name	D-II weight (%)	D-II name
CD5x	1	951	s.w.	73.66	5s	65.04	2Z,S,Z,S
	2	413	s.w.	25.27	4s,-	12.60	2Z,S,-,S
CD5	1	42	s.w.	99.80	5s	59.28	2Z,S,Z,S
	2	451	s.w.	0.20	4s,-	0.20	2Z,S,-,S
CD6	1	160	s.w.	89.95	6s	18.24	2Z,S,2Z,S
	2	91	s.w.	8.20	4s,+,-	0.29	Z,S,Z,+,-,S
CD7	1	751	s.w.	64.29	7s	19.41	2Z,S,Z,S,Z,S
	2	1013	s.w.	11.80	6s,-	2.05	2Z,2S,-,Z,S
	3	867	s.w.	8.00	5s,+,-	4.39	2Z,S,2Z,+,-
CD8	1	257	s.w.	63.41	8s	9.95	Z,S,Z,S,Z,S,Z,S
	2	963	s.w.	8.39	6s,+,-	1.27	2Z,a,-,Z,S,Z,S
	3	657	s.w.	4.59	6s,a,-	4.78	2Z,2S,2Z,a,-
	4	974	s.w.	4.49	7s,-	0.39	3Z,2S,-,2S
CD14	1	720	s.w.	8.10	13s,-	0.02	2Z,S,Z,S,Z,2S,2Z,S,-,2S
	2	1320	s.w.	5.25	14s	0.05	3Z,S,Z,S,Z,S,Z,S,Z,S,Z,S
	3	3825	rand.	0.02	7s,-,5s,a	0.02	3Z,-,S,Z,2S,Z,a,Z,3S
CD21	1	1846	s.w.	0.59	c.n.	0.02	c.n.
	2	2124	rand.	0.02	c.n.	0.02	c.n.
	3	1916	rand.	0.02	c.n.	0.02	c.n.
CD26	1	2712	rand.	0.02	c.n.	0.02	c.n.
	2	2831	rand.	0.02	c.n.	0.02	c.n.
	3	1266	rand.	0.02	c.n.	0.02	c.n.
CD28	1	2163	rand.	0.02	c.n.	0.02	c.n.
	2	3445	rand.	0.02	c.n.	0.02	c.n.
	3	3007	rand.	0.02	c.n.	0.02	c.n.

Figure 5.50: Table of structure selection for MD calculations. Selection criteria were *statistical weight* "s.w." and *random* "rand". Conformations labelled "c.n." represent those ones selected from *conformational noise.*

The table shows the number of conformations selected for every cyclodextrin, the number of every conformation in the Markov chain, the selection criterion, and the statistical weights and names according to D-I and D-II.



Johannes Brahms
Symphony No. 4 in e minor op. 98
(1885)
IV. Movement. Chaconne
(Passacaglia theme from J.S. Bach's Cantata BWV 150
"Nach dir, Herr, verlanget mich")

"I have yet to see any problem, however complicated,
which, when you looked at it the right way,
did not become still more complicated"

Paul Alderson

6 RESULTS III: MOLECULAR DYNAMICS "IN VACUO"

6.1 MOLECULAR DYNAMICS

Both this chapter and the next one contain the second part of the research; the Molecular Dynamics (MD) calculations. Within this frame, two objectives were established:

- Test the efficiency of MD methodology in combination with Trajectory Overlapping Analysis to prove/false MD dependence on starting conformation.
- Since MD models the *real* behaviour of molecules in vacuo/solvent; explore the Conformational Space of the CDs[195] selected as benchmark and do a comparative study of their conformational properties.

Molecular Dynamics (see appendix: chapter IX methodology) has been applied over the entire group of CDs:

- Macrorings including the set of RESP-charges calculated by Dr. Iván Beà and also used by Dr. Itziar Maestre and Dr. Miguel de Federico:
  - ❖ Small cyclodextrins: CD5, CD6, CD7 and CD8.
  - ❖ Large cyclodextrins: CD14, CD21, CD26 and CD28.
- Macrorings including the set of RESP-charges calculated by Dr. Javier Pérez:
  - ❖ Small cyclodextrin CD5x (where "x" stands for the new set of charges).

For every cyclodextrin, several MD trajectories –starting from different conformations- have been obtained in order to measure the overlapping area in conformational space. With the purposes of evaluating the influence of the presence/absence of solvent and the polarity in MD calculations, the series of MD were carried out under different environments. The current chapter, however, focuses in the gas phase calculations:

- Gas Phase (*current chapter*): CD5, CD6, CD7, CD8, CD14, CD21, CD26, CD28 and CD5x.
- Water (*next chapter*): CD8, CD14, CD21, CD26, CD28.
- Benzene (*next chapter*): CD26.

[195] (a) Ivanov, P.M.; Jaime C.; *J. Phys. Chem. B*. **2004**. *108(20)*. 6261-6274. (b) Gotsev, M.G.; Ivanov, P.M.; Jaime, C.; *Chirality*. **2007**. *19(3)*. 203-213. (c) Perez-Miron, J.; Jaime, C.; Ivanov, P.M.; *Chirality*. **2008**. *20(10)*. 1127-1133.

Descriptors I and II have been calculated for the whole sets of MD trajectories although only the second one has been extensively used in the present Thesis. Complete information regarding descriptors I and II, Markov matrices, statistics and chain of conformations not included in this chapter can be found in the DVD attached to the back cover.

6.2 Molecular Dynamics "in vacuo".

The series of MD calculations in gas phase were carried out following the schema in [Fig. 6.1]. The full calculation included two steps:

- The first step is the *heating slope* and *equilibration* process and includes *"in vacuo"* conditions. Its length is 300 picoseconds, timestep of 1 femtosecond, cutoff of 12 angstrom, no box, no PBC, 298 K in the equilibrium and external constraints to avoid chair-to-boat conformational interchanges in glucoses. Coupling constants are modulated along the simulation to avoid *blow-up* terminations.
- The second step is the *sampling* process including also *"in vacuo"* conditions. Its length is 10,000 picoseconds, timestep of 1 femtosecond, cutoff of 12 angstrom, no box, no PBC, 298 K, bath thermal coupling of 0.5 picoseconds and sampling frequency of 0.1 snapshots per picosecond –the trajectory is stored in this step for further analysis-. External constraints –to avoid chair-to-boat conformational interchanges in glucoses- are removed and coupling constants are kept unchanged along the simulation.

Detailed information regarding gas phase MD conditions and further computational aspects can be found in the DVD: annex 10.1.1.4 and its sub-annexes (chapter X AMBER 7 files).

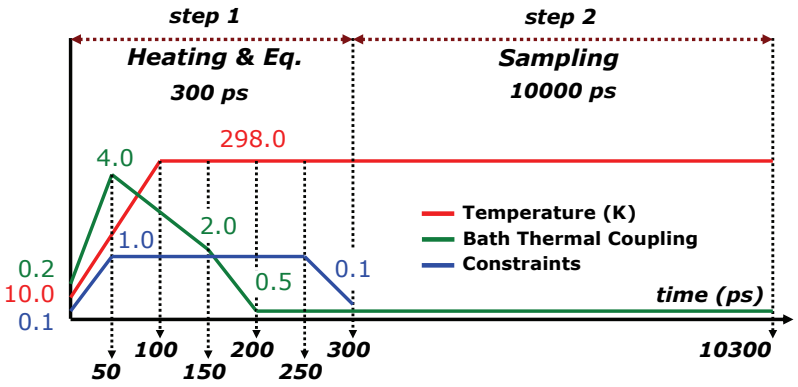


Figure 6.1: Schematic depiction of the 2-step “in vacuo” Molecular Dynamics protocol. Coloured lines, although not scaled, represent the dynamically coupled control parameters.

6.2.1 CD5

Two Molecular Dynamics have been done; one starting from [**2Z,S,Z,S**] and the other from [**2Z,S,-,S**]. Both conformations were selected according to statistical weight criteria.

6.2.1.1 Population Analysis

The results can be found in [Fig. 6.2]. Some important conclusions can be derived from the table:

- The total number of conformations found is only 8, which is a small number.
- The trajectories t1, t2 and the combined trajectory –the average- share in common the majority of their conformations.
- The combined trajectory has almost the same number of conformations than t1 and t2.
- The most highly rated conformations in t1, t2 and the combined trajectory are: [**2Z,S,Z,S**], [**Z,2S,Z,S**] and [**4Z,S**].
- The most populated conformations detected by Molecular Dynamics in the combined trajectory are coincident with the highly populated ones found in the Simulated Annealing Conformational Search.
- While the starting conformation in t1 is also coincident with the most populated conformation, the situation is completely different in t2, where the starting

conformation is not even detected; this meaning that MD calculations do not necessarily get stuck in the starting point.

Trajectory	1		2		average	
CD5	name D-II	weight (%)	name D-II	weight (%)	name D-II	weight (%)
Starting Conf.	2Z,S,Z,S	(-)	2Z,S,-,S	(-)		
MD Conf.	2Z,S,Z,S	64.10	Z,2S,Z,S	51.00	2Z,S,Z,S	55.70
	4Z,S	18.80	2Z,S,Z,S	47.30	Z,2S,Z,S	31.35
	Z,2S,Z,S	11.70	4Z,S	1.20	4Z,S	10.00
	3Z,2S	3.40	3Z,2S	0.30	3Z,2S	1.85
	5Z	1.80	Z,4S	0.10	5Z	0.90
	2Z,S,-,S	0.10	2Z,3S	0.10	2Z,3S	0.10
	2Z,3S	0.10			Z,4S	0.05
					2Z,S,-,S	0.05
TOTAL (%)	7	100.00	6	100.00	8	100.00

Figure 6.2: Conformations from CD5 "in vacuo" MD trajectories 1, 2, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.1.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* –or *static lambda index*- was already defined in chapter IV. Now, on the basis of the *lambda index*, we define the *accumulative lambda index* –or *dynamic lambda index*- as the collection of all the *static lambda indexes* calculated every time the number of conformations in every trajectory is incremented by one. The graphic of the set of *static lambda indexes* vs. time is what we call the *accumulative* or *dynamic lambda index* and is also as a quantitative way of estimating the trajectory overlapping degree as a function of time.

The total number of conformations in t1, t2 and the combined trajectory [Fig. 6.3a] helps calculating the *accumulative lambda index* [Fig. 6.3b].

- Since t1 –in blue- and the combined trajectory –in red- evolve similarly [Fig. 6.3a], it is worth to say that they cover almost the same conformational space and therefore t1 is a reasonable source of information by itself.
- Although t1 is a good option, t1 and t2 separately cannot fully explain the whole system. Nevertheless, the resulting information from the combined trajectory is rather precise in describing the conformational space.
- Trajectories t1 and t2, at the end, have in common almost 77% of their conformational spaces which is higher than 3 parts in 4 [Fig. 6.3b].

Unfortunately, it still seems to be quite a big distance to 100% overlapping. The truth is that lambda index becomes stricter as the total number of conformations computed is smaller; 7 in t1, 6 in t2 and 8 in average.

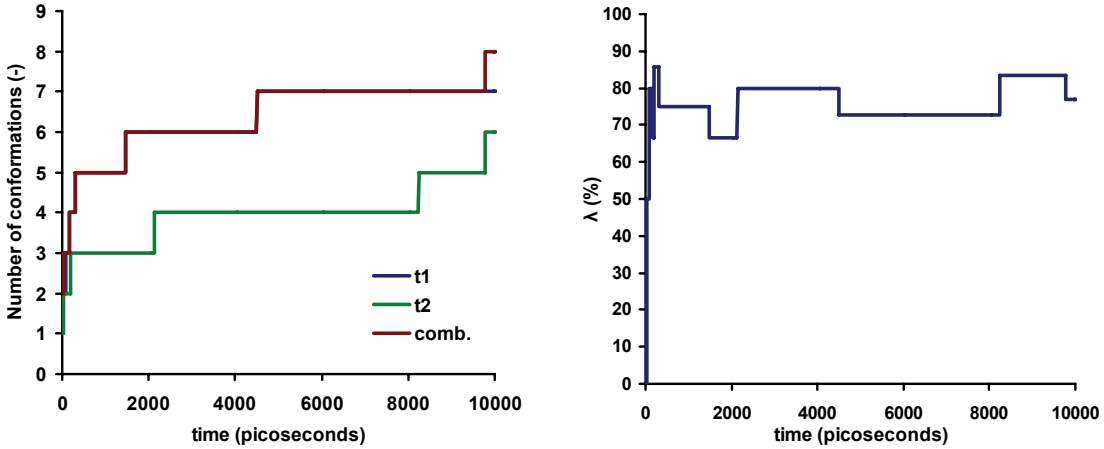


Figure 6.3: (a) Number of conformations in *t1*, *t2* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD5.

In fact, there is only one conformation in t2 that was not detected in t1: [**Z,4S**], and two conformations in t1 that were not detected in t2; [**5Z**] and [**2Z,S,-,S**]. This result is also confirmed checking the omega index table [Fig. 6.4] where those three conformations have omega values of 0%.

6.2.1.3 Overlapping Analysis II: Omega Index

The omega index was already defined in Chapter IV. It quantifies, once it is known that there is overlapping in a group of trajectories, which of the conformations within the trajectories are involved in the overlapping effect and which of them are not. This index is calculated for the conformation exhibiting the higher ratio and then it is said that the other trajectories overlap the higher populated one. On the grounds of these assumptions it can be said that:

- The most important conformations of CD5: [**2Z,S,Z,S**] with 55.70%, [**Z,2S,Z,S**] with 31.35% and [**4Z,S**] with 10.00% experiment overlapping. The first one –predominant in t1- is 77.79% overlapped by t2. The second one –predominant in

t2- is 22.94% overlapped by t1. And the third one –predominant in t1- is 6.38% overlapped by t2.

- The conformations with omega values of 0% are those that have been detected either in t1 or t2, but not in both the trajectories at the same time. These conformations are located in those parts of the trajectories that do not overlap.

name D-II	t1	t2	$<{}^{i}p_{i}>$	${}^{i}p_{i(max)}$	${}^{i}p_{i(max)}/r$	${}^{i}v_{j}$	${}^{i}d_{j}$	${}^{i}\omega_{i}(\%)$
2Z,S,Z,S	64.10	47.30	**55.70**	64.10	32.05	23.65	32.05	**73.79**
Z,2S,Z,S	11.70	51.00	**31.35**	51.00	25.50	5.85	25.50	**22.94**
4Z,S	18.80	1.20	**10.00**	18.80	9.40	0.60	9.40	**6.38**
3Z,2S	3.40	0.30	**1.85**	3.40	1.70	0.15	1.70	**8.82**
5Z	1.80	0.00	**0.90**	1.80	0.90	0.00	0.90	**0.00**
2Z,3S	0.10	0.10	**0.10**	0.10	0.05	0.05	0.05	**100.00**
Z,4S	0.00	0.10	**0.05**	0.10	0.05	0.00	0.05	**0.00**
2Z,S,-,S	0.10	0.00	**0.05**	0.10	0.05	0.00	0.05	**0.00**
TOTAL (%)			**100.00**					

Figure 6.4: Omega index –individual conformations overlapping ratio- for CD5.

6.2.1.4 Average Structures

As was said in chapter V, the average structure of a set of conformations is an *artificial* structure which mostly represents their overall behaviour. When certain conformations weight more than others, their contribution to the average structure is noticeable and, in an important degree, the average structure takes a lot from them. Therefore, the *average structure* is probably the most similar conformation to the *main conformation*.

Similarly to what was done in chapter V, the images in [Fig. 6.5] and other ones in further sections describing *small cyclodextrins* have been created with ***Mercury molecular viewer***[196]. A PDB file containing the average structure was previously computed with AMBER 7 module ptraj processing the MD trajectory file; t1 and t2. The command file can be consulted in the DVD: appendix 10.3.1.2 (chapter X Amber 7 files).

[196] (a) **Mercury - *Crystal Structure Visualisation and Exploration Made Easy***. Copyright © 2004-2009 The Cambridge Crystallographic Data Centre. 12 Union Road, Cambridge, CB2 1EZ, UK, +44 1223 336408. *Registered in England No.2155347. Registered Charity No.800579.*
(b) http://www.ccdc.cam.ac.uk/mercury/ (c) http://www.ccdc.cam.ac.uk/free_services/mercury/

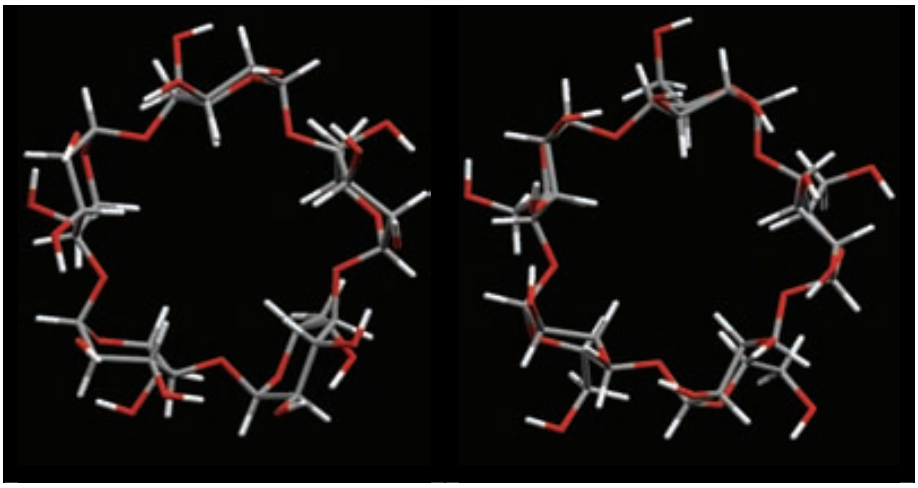


Figure 6.5: Average conformations from CD5 “in vacuo” MD trajectories t1 and t2. The point of view of the images is the face bearing the secondary hydroxyls.

The images support the results obtained from the descriptor analysis:

- The average structures from t1 and t2 are viable molecules.
- Rigidity is important in these macrocycles and they seem to keep a cavity available for host-guest phenomena. However, Solvent Accessible Surface Area calculations show that the cavity is really small, better promoting associative complexes than inclusive ones.
- Both t1 and t2 arrange their primary hydroxyls in a clockwise chain of hydrogen bonds. Otherwise, the average structure from t1 arranges its secondary hydroxyls in a counterclockwise chain and the one from t2 arranges them in a clockwise chain.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms cannot be detected by D-I or D-II; however, the overall behaviour of the macrocycle is well described.
- The long chains of hydrogen bonds detected in gas phase calculations involving primary and secondary hydroxyls are important to stabilize the macrocycle and therefore those conformations that allow them are the ones preferred.

6.2.2 CD5x

This cyclodextrin was also studied when the Simulated Annealing Conformational Search was carried out. As was then said, CD5x is related to CD5 only in part: both of them are similar in all aspects –the same force field and identical number of glucoses- but they differ in the set of charges employed. While CD5 implements the set of charges calculated by Dr. Beà, CD5x uses a different set developed by Dr. Pérez. The point of this change is evaluating the effect of the set of charges in the MD calculations as was previously done with the SA calculations.

Therefore, as in the case of CD5, two Molecular Dynamics have been done; one starting from [**2Z,S,Z,S**] and the other from [**2Z,S,-,S**]. Both conformations were selected according to statistical weight criteria and –also important- were also the starting conformations employed in the CD5 series of MD calculations. Then, under these conditions, comparing the results of CD5 and CD5x is fully allowed.

6.2.2.1 Population Analysis

The results of the MD series can be found in [Fig. 6.7]. Some important conclusions derived from the table and further comparisons with CD5 are included hereafter:

- The total number of conformations found is 17 which –although not a very big number- is more than twice the number of conformations found in CD5.
- The trajectories t2 and the combined trajectory share in common most of their conformations. However, t1 has only half the number of conformations of t2. This behaviour is significantly different from that observed in CD5, where t1 and t2 were mostly similar in number and type of conformations.
- The most highly rated conformations in t1, t2 and the combined trajectory are: [**2Z,S,Z,S**], [**4Z,S**] and [**Z,2S,Z,S**]. They are the very same conformations –in number and type- found as the highly populated ones in CD5. However, there are two differences. On the one hand, the 2nd and 3rd positions are interchanged in CD5x and CD5. On the other hand, the ratios have changed [Fig. 6.6],

enriching conformation [**2Z,S,Z,S**] in CD5x, which means that the new set of charges favours this particular conformation at the expense of others.

Conformation	MD statistical weight (%)	
	CD5	CD5x
2Z,S,Z,S	55.70	64.10
Z,2S,Z,S	31.35	9.65
4Z,S	10.00	11.25
TOTAL	**97.05**	**85.00**

Figure 6.6: Statistical weights of the three most important conformations found in CD5 and CD5x "in vacuo" MD combined trajectories.

- The most populated conformations detected by Molecular Dynamics in the combined trajectory are coincident with the highly populated ones found in the CD5 series of MD calculations and also with the Simulated Annealing Conformational Search.
- While the starting conformation in t1 is also coincident with the most populated conformation, the situation is different in t2, where the starting conformation is detected in the 5th position; this means, again, that MD calculations not necessarily get stuck in the starting conformations.

Trajectory	1		2		average	
CD5x	name D-II	weight (%)	name D-II	weight (%)	name D-II	weight (%)
Starting Conf.	2Z,S,Z,S	(-)	2Z,S,-,S	(-)		
MD Conf.	2Z,S,Z,S	74.30	2Z,S,Z,S	53.90	2Z,S,Z,S	64.10
	4Z,S	14.40	Z,2S,Z,S	11.20	4Z,S	11.25
	Z,2S,Z,S	8.10	4Z,S	8.10	Z,2S,Z,S	9.65
	5Z	1.60	Z,2S,-,S	6.00	Z,2S,-,S	3.00
	3Z,2S	1.50	2Z,S,-,S	5.50	2Z,S,-,S	2.75
	2Z,3S	0.10	Z,S,-,2S	5.10	Z,S,-,2S	2.55
			2S,-,S,-	4.70	2S,-,S,-	2.35
			3Z,2S	2.60	3Z,2S	2.05
			5Z	1.00	5Z	1.30
			3Z,S,-	0.60	3Z,S,-	0.30
			2Z,2S,-	0.50	2Z,2S,-	0.25
			Z,S,Z,-,S	0.20	Z,S,Z,-,S	0.10
			Z,S,-,S,-	0.20	Z,S,-,S,-	0.10
			2Z,-,2S	0.20	2Z,-,2S	0.10
			Z,S,Z,S,-	0.10	Z,S,Z,S,-	0.05
			Z,-,S,-,S	0.10	Z,-,S,-,S	0.05
					2Z,3S	0.05
TOTAL (%)	6	100.00	16	100.00	17	100.00

Figure 6.7: Conformations from CD5x "in vacuo" MD trajectories 1, 2, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.2.2 Overlapping Analysis I: Accumulative Lambda Index

The *accumulative lambda index* [Fig. 6.8b] is calculated from the total number of conformations in t1, t2 and the combined trajectory [Fig. 6.8a].

- Since t2 –in green- and the combined trajectory –in red- evolve similarly [Fig. 6.8a], it is reasonable to say that they cover almost the same conformational space and therefore t2 is a reasonable source of information by itself. In this sense, although t2 fully saturates, the area explored in conformational space is considerably smaller. This situation is different in CD5, where the two of them were almost equally efficient in the exploration.
- Although t2 is a good option, t1 and t2 separately cannot fully explain the whole system. Nevertheless, the resulting information from the combined trajectory is rather precise in describing the conformational space. In this sense, this result is similar to that of CD5.
- Trajectories t1 and t2, at the end, have in common nearly 45% of their conformational spaces which is slightly less than 1 part in 2 [Fig. 6.8b]. This rate is considerably worse than that obtained for CD5 –where the overlapping was close to 77%-, particularly considering the fact that this is the smaller cyclodextrin and should have been extremely well sampled. However, this result must be correctly understood: detailed examination of both graphics in [Fig. 6.8] shows that the plot of the number of conformations versus time –the individual t1, t2, and the combined trajectories- and the accumulative lambda index saturate over 5000 picoseconds. Therefore, it is not so important that the individual trajectories do not fully overlap while the combined trajectory saturates, which means that all conformations –at least the most important ones- have been detected.

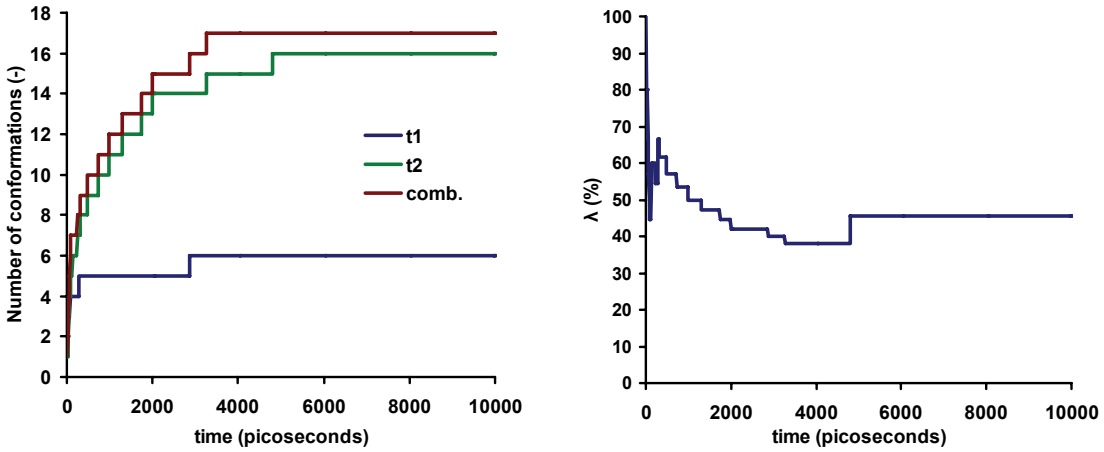


Figure 6.8: (a) Number of conformations in *t1*, *t2* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD5x.

Comparing both trajectories –as was done with CD5- it is found that there is only one conformation in t1 that was not detected in t2: [**2Z,3S**], with a statistical weight of 0.10% in t1 and 0.05% in the combined trajectory. Meanwhile, there are eleven conformations in t2 that were not detected in t1; [**Z,2S,-,S**], [**2Z,S,-,S**], [**Z,S,-,2S**], [**2S,-,S,-**], [**3Z,S,-**], [**2Z,2S,-**], [**Z,S,Z,-,S**], [**Z,S,-,S,-**], [**2Z,-,2S**], [**Z,S,Z,S,-**] and [**Z,-,S,-,S**]. This group has a statistical weight of 23.20% in t2 and a total weight of 11.60% in the combined trajectory. This result is also confirmed checking the omega index table [Fig. 6.9] where those conformations have omega values of 0%.

6.2.2.3 Overlapping Analysis II: Omega Index

The omega indexes have been calculated for the whole set of CD5x conformations and some remarks can be mentioned:

- The most important conformations of CD5x: [**2Z,S,Z,S**] with 64.10%, [**4Z,S**] with 11.25% and [**Z,2S,Z,S**] with 9.65% experiment overlapping. The first one –predominant in t1- is 72.54% overlapped by t2. The second one –predominant also in t1- is 56.25% overlapped by t2. And the third one –predominant in t2- is 72.32% overlapped by t1.

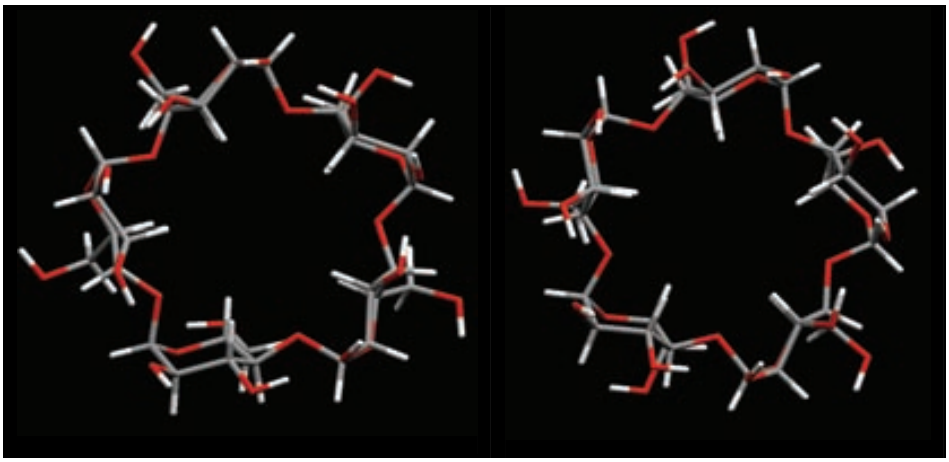
- The conformations with omega values of 0% are those that have been detected either in t1 or t2, but not in both trajectories at the same time. These conformations are located in those parts of the trajectories that do not overlap.

name D-II	t1	t2	$<^{i}p_{i}>$	$^{i}p_{i(max)}$	$^{i}p_{i(max)}/r$	$^{i}\nu_{i}$	$^{i}d_{i}$	$^{i}\omega_{i}(\%)$
2Z,S,Z,S	74.30	53.90	**64.10**	74.30	37.15	26.95	37.15	**72.54**
4Z,S	14.40	8.10	**11.25**	14.40	7.20	4.05	7.20	**56.25**
Z,2S,Z,S	8.10	11.20	**9.65**	11.20	5.60	4.05	5.60	**72.32**
Z,2S,-,S	0.00	6.00	**3.00**	6.00	3.00	0.00	3.00	**0.00**
2Z,S,-,S	0.00	5.50	**2.75**	5.50	2.75	0.00	2.75	**0.00**
Z,S,-,2S	0.00	5.10	**2.55**	5.10	2.55	0.00	2.55	**0.00**
2S,-,S,-	0.00	4.70	**2.35**	4.70	2.35	0.00	2.35	**0.00**
3Z,2S	1.50	2.60	**2.05**	2.60	1.30	0.75	1.30	**57.69**
5Z	1.60	1.00	**1.30**	1.60	0.80	0.50	0.80	**62.50**
3Z,S,-	0.00	0.60	**0.30**	0.60	0.30	0.00	0.30	**0.00**
2Z,2S,-	0.00	0.50	**0.25**	0.50	0.25	0.00	0.25	**0.00**
Z,S,Z,-,S	0.00	0.20	**0.10**	0.20	0.10	0.00	0.10	**0.00**
Z,S,-,S,-	0.00	0.20	**0.10**	0.20	0.10	0.00	0.10	**0.00**
2Z,-,2S	0.00	0.20	**0.10**	0.20	0.10	0.00	0.10	**0.00**
Z,S,Z,S,-	0.00	0.10	**0.05**	0.10	0.05	0.00	0.05	**0.00**
Z,-,S,-,S	0.00	0.10	**0.05**	0.10	0.05	0.00	0.05	**0.00**
2Z,3S	0.10	0.00	**0.05**	0.10	0.05	0.00	0.05	**0.00**
TOTAL (%)			**100.00**					

Figure 6.9: Omega index –individual conformations overlapping ratio- for CD5x.

6.2.2.4 Average Structures

Since CD5x is again a small cyclodextrin the ordinary view is employed here.



t1 **t2**

Figure 6.10: Average conformations from CD5x "in vacuo" MD trajectories t1 and t2. The point of view of the images is the face bearing the secondary hydroxyls.

The images in [Fig. 6.10] support the results obtained from the descriptor analysis, which is also what happened with the small CDs bearing the other set of charges.

- The average structures from t1 and t2 are viable molecules. The conformations, however, are slightly distorted in comparison to those of CD5, probably as a consequence of the subtle increase in flexibility.
- Similarly to CD5, rigidity is important in these macrocycles maintaining a cavity available for host-guest phenomena; nevertheless, Solvent Accessible Surface Area calculations show that the cavity is really small, and therefore associative complexes would be favoured.
- Both average structures from t1 and t2 arrange their primary hydroxyls in a clockwise chain of hydrogen bonds. Likewise, t1 and t2 secondary hydroxyls are both oriented in a counterclockwise chain.
- As was said in the case of CD5, the fact that both the clockwise and the counterclockwise directions were detected in CD5, while only the counterclockwise direction was detected in CD5x, clearly states that the directionality involving hydrogen bonds between hydroxyls and oxygen atoms cannot be detected by D-I or D-II. The overall behaviour of the macrocycle is, however, well described.
- The long chains of hydrogen bonds detected in gas phase calculations involving primary and secondary hydroxyls are important to stabilize the macrocycle and therefore those conformations that allow them are the ones preferred.

6.2.3 CD6 (α-CD)

Again, two Molecular Dynamics have been done; the first one starting from [**2Z,S,2Z,S**] and the second one from [**Z,S,Z,+,-,S**]. Both conformations were selected according to statistical weight criteria.

6.2.3.1 Population Analysis

The results can be found in [Fig. 6.11]. Some important conclusions can be derived from the table:

- The total number of conformations found is only 14, which is still a small number for such a big molecule.
- The trajectories t1, t2 and the combined trajectory –the average- share in common many important conformations.
- The combined trajectory has almost the same number of conformations than t2. However, the number of conformations in t1 is lower.
- The most highly rated conformations in t1, t2 and the combined trajectory are: [**Z,S,Z,S,Z,S**], [**2Z,S,2Z,S**], [**2Z,S,Z,2S**], [**3Z,S,Z,S**] and [**Z,2S,Z,2S**].
- The most populated conformations detected by Molecular Dynamics in the combined trajectory are coincident with the highly populated ones found in the Simulated Annealing Conformational Search.
- As happened in CD5; while the starting conformation in t1 is also coincident with the most populated conformation, the situation is completely different in t2, where the starting conformation is not even detected; this means that MD calculations do not necessarily get stuck in the starting conformations.

Trajectory	1		2		average	
CD6	name D-II	weight (%)	name D-II	weight (%)	name D-II	weight (%)
Starting Conf.	2Z,S,2Z,S	(-)	Z,S,Z,+,-,S	(-)		
MD Conf.	2Z,S,2Z,S	38.10	Z,S,Z,S,Z,S	55.70	Z,S,Z,S,Z,S	36.70
	Z,S,Z,S,Z,S	17.70	Z,2S,Z,2S	15.70	2Z,S,2Z,S	22.15
	3Z,S,Z,S	16.30	2Z,S,Z,2S	13.80	2Z,S,Z,2S	14.30
	2Z,S,Z,2S	14.80	2Z,S,2Z,S	6.20	3Z,S,Z,S	9.45
	2Z,2S,Z,S	8.50	2Z,2S,Z,S	3.90	Z,2S,Z,2S	8.70
	5Z,S	2.30	3Z,S,Z,S	2.60	2Z,2S,Z,S	6.20
	Z,2S,Z,2S	1.70	Z,2S,-,2S	0.80	5Z,S	1.15
	4Z,2S	0.50	Z,3S,Z,S	0.60	Z,2S,-,2S	0.40
	2Z,2S,-,S	0.10	Z,S,Z,S,-,S	0.20	Z,3S,Z,S	0.30
			2Z,S,-,2S	0.20	4Z,2S	0.30
			4Z,2S	0.10	Z,S,Z,S,-,S	0.10
			3Z,S,-,S	0.10	2Z,S,-,2S	0.10
			2Z,2S,-,S	0.10	2Z,2S,-,S	0.10
					3Z,S,-,S	0.05
TOTAL (%)	9	100.00	13	100.00	14	100.00

Figure 6.11: Conformations from CD6 "in vacuo" MD trajectories 1, 2, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.3.2 Overlapping Analysis I: Accumulative Lambda Index

Considering the total number of conformations in t1, t2 and the combined trajectory [Fig. 6.12a] the *accumulative lambda index* [Fig. 6.12b] is calculated:

- Both trajectories t1 and t2 evolve in time towards better explorations of the conformational space, although t2 –in green- seems to find new conformations over 6000 picoseconds while t1 gets stuck in a certain area [Fig. 6.12a]. Over 8000 picoseconds, the tendency in t1, t2 and the combined trajectory suggests that the system has reached a dynamic equilibrium.
- Although t1 and t2 separately cannot explain the whole system, the combined information is still helpful in describing the conformational space.
- Trajectories t1 and t2 have in common almost 73% of their conformational spaces which is nearly 3 parts in 4.

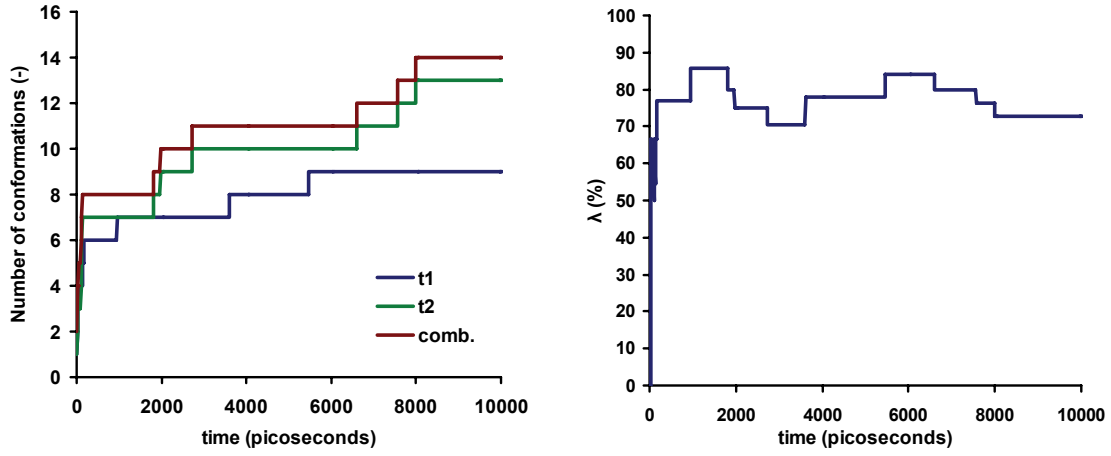


Figure 6.12: (a) Number of conformations in *t1*, *t2* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD6.

Under detailed examination, it can be seen that t2 did a wider exploration of the conformational space of CD6, visiting more conformations than t1: There is only one conformation in t1 that was not detected in t2; [**5Z,S**], meanwhile, there are five conformations in t2 that were not detected in t1; [**Z,2S,-,2S**], [**Z,3S,Z,S**], [**Z,S,Z,S,-,S**], [**2Z,S,-,2S**] and [**3Z,S,-,S**]. This result is also confirmed checking the omega index table [Fig. 6.13] where those six conformations have omega values of 0% in either t1 or t2.

6.2.3.3 Overlapping Analysis II: Omega Index

The omega indexes have been calculated for the whole set of CD6 conformations and some remarks can be mentioned:

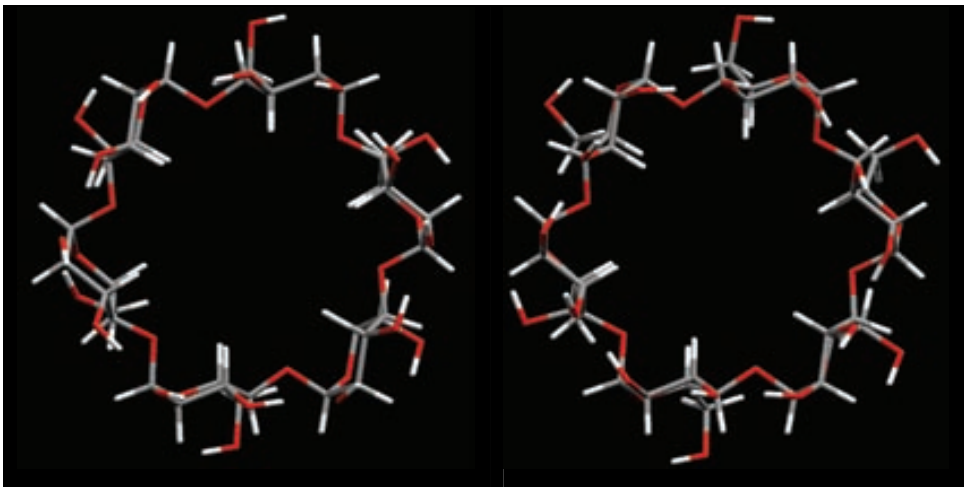
- The most important conformations of CD6 are: [**Z,S,Z,S,Z,S**] with 36.70%, [**2Z,S,2Z,S**] with 22.15%, [**2Z,S,Z,2S**] with 14.30%, [**3Z,S,Z,S**] with 9.45% and [**Z,2S,Z,2S**] with 8.70% experiment overlapping. The first one –predominant in t2- is 31.78% overlapped by t1. The second one –predominant in t1- is 16.27% overlapped by t2. The third one –predominant in t1- is 93.24% overlapped by t2. The fourth one –predominant in t1- is 15.95% overlapped by t2. And the fifth one –predominant in t2- is 10.83% overlapped by t1.
- The conformations with omega values of 0% are those that have been detected either in t1 or t2, but not in both trajectories at the same time.

name D-II	t1	t2	$<{}^{i}p_j>$	${}^{i}p_{j(max)}$	${}^{i}p_{j(max)}/r$	${}^{i}\nu_j$	${}^{i}d_j$	${}^{i}\omega_j(\%)$
Z,S,Z,S,Z,S	17.70	55.70	**36.70**	55.70	27.85	8.85	27.85	**31.78**
2Z,S,2Z,S	38.10	6.20	**22.15**	38.10	19.05	3.10	19.05	**16.27**
2Z,S,Z,2S	14.80	13.80	**14.30**	14.80	7.40	6.90	7.40	**93.24**
3Z,S,Z,S	16.30	2.60	**9.45**	16.30	8.15	1.30	8.15	**15.95**
Z,2S,Z,2S	1.70	15.70	**8.70**	15.70	7.85	0.85	7.85	**10.83**
2Z,2S,Z,S	8.50	3.90	**6.20**	8.50	4.25	1.95	4.25	**45.88**
5Z,S	2.30	0.00	**1.15**	2.30	1.15	0.00	1.15	**0.00**
Z,2S,-,2S	0.00	0.80	**0.40**	0.80	0.40	0.00	0.40	**0.00**
Z,3S,Z,S	0.00	0.60	**0.30**	0.60	0.30	0.00	0.30	**0.00**
4Z,2S	0.50	0.10	**0.30**	0.50	0.25	0.05	0.25	**20.00**
Z,S,Z,S,-,S	0.00	0.20	**0.10**	0.20	0.10	0.00	0.10	**0.00**
2Z,S,-,2S	0.00	0.20	**0.10**	0.20	0.10	0.00	0.10	**0.00**
2Z,2S,-,S	0.10	0.10	**0.10**	0.10	0.05	0.05	0.05	**100.00**
3Z,S,-,S	0.00	0.10	**0.05**	0.10	0.05	0.00	0.05	**0.00**
TOTAL (%)			**100.00**					

Figure 6.13: Omega index –individual conformations overlapping ratio- for CD6.

6.2.3.4 Average Structures

The *average structure* is probably the most similar conformation to the *main conformation* and therefore it helps guessing the overall molecular behaviour in time.



t1 t2

Figure 6.14: Average conformations from CD6 “in vacuo” MD trajectories t1 and t2. The point of view of the images is the face bearing the secondary hydroxyls.

The images [Fig. 6.14] support the results obtained from the descriptor analysis:

- The average structures from t1 and t2 are viable molecules.
- Rigidity –although slightly more flexible than CD5- is important in these macrocycles keeping a cavity available for host-guest phenomena. Solvent Accessible Surface Area calculations show that the cavity is bigger than that of CD5 allowing both associative and inclusive complexes.
- Both t1 and t2 arrange their primary hydroxyls in a clockwise chain of hydrogen bonds. On the other hand, the average structure from t1 arranges its secondary hydroxyls in a counterclockwise chain and the one from t2 arranges them in a clockwise chain.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms cannot be detected by D-I or D-II; however, the overall behaviour of the macrocycle is well described.
- The long chains of hydrogen bonds detected in gas phase calculations involving primary and secondary hydroxyls are important to stabilize the macrocycle and therefore those conformations that allow them are the ones preferred.

6.2.4 CD7 (β-CD)

The set of Molecular Dynamics carried out for CD7 is different because three different starting conformations were selected instead of the two conformations used with CD5 and CD6. The increase in CD7 flexibility strongly recommended this update in the protocol. The group of the starting conformations includes: the first one [**2Z,S,Z,S,Z,S**]; the second one [**2Z,2S,-,Z,S**] and the third one [**2Z,S,2Z,+,-**]. All of them were selected according to statistical weight criteria.

6.2.4.1 Population Analysis

The most representative results can be found in [Fig. 6.15]. Just 16 conformations have been selected and represent, in all cases, over 99% of the system. Some important conclusions can be derived from the table. They are explained hereafter:

- The total number of conformations found is 31, which is still fortunately an easy-handling number for a cyclodextrin.
- The trajectories t1, t2, t3 and the combined trajectory –the average- share in common some of the most important conformations, at least the first three ones.
- The total number of conformations in t1, t2, t3 and the combined trajectory grows as the number of glucoses in the cyclodextrin is increased. In this sense, t1 and t2 present a similar number of conformations, followed by t3 –which is slightly bigger-, the total number of conformations in the combined trajectory being considerably bigger in comparison to the individual trajectories.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**2Z,S,Z,S,Z,S**], [**2Z,2S,2Z,S**] and [**3Z,S,2Z,S**].
- Some of the most populated conformations detected by Molecular Dynamics in the combined trajectory are coincident with the highly populated ones found in the SA Conformational Search. However, conformation [**3Z,S,2Z,S**] –the one rated 3rd during MD calculation- was found in the 12th position –probability 1.46%- meaning that rankings change for those conformations beyond the group of the highly populated ones.

- The starting conformation in t1 is also the most populated conformation. The starting conformation in t2 is not detected in the whole trajectory. And the starting conformation in t3 is rated the 8th in the trajectory population analysis. The variety of possibilities seems to suggest that MD calculations do not necessarily get stuck in the starting conformations.

Trajectory	1		2		3		average	
CD7	name D-II	w (%)	name D-II	w (%)	name D-II	w (%)	name D-II	w (%)
Starting Conf.	2Z,S,Z,S,Z,S	(-)	2Z,2S,-,Z,S	(-)	2Z,S,2Z,+,-	(-)		
MD Conf.	2Z,S,Z,S,Z,S	48.70	2Z,S,Z,S,Z,S	51.20	2Z,S,Z,S,Z,S	40.80	2Z,S,Z,S,Z,S	46.90
	2Z,2S,2Z,S	19.90	2Z,2S,2Z,S	22.90	2Z,2S,2Z,S	20.70	2Z,2S,2Z,S	21.17
	3Z,S,2Z,S	19.50	3Z,S,2Z,S	15.90	3Z,S,2Z,S	17.00	3Z,S,2Z,S	17.47
	Z,2S,Z,S,Z,S	5.10	Z,2S,Z,S,Z,S	3.70	2Z,a,-,Z,2S	5.90	Z,2S,Z,S,Z,S	3.83
	2Z,2S,Z,2S	2.60	2Z,2S,Z,2S	2.40	Z,2S,Z,S,Z,S	2.70	2Z,2S,Z,2S	2.40
	4Z,S,Z,S	1.20	4Z,S,Z,S	1.80	2Z,S,2Z,a,-	2.70	2Z,a,-,Z,2S	1.97
	3Z,S,Z,2S	0.90	3Z,2S,Z,S	0.60	2Z,2S,Z,2S	2.20	4Z,S,Z,S	1.30
	3Z,2S,Z,S	0.50	3Z,S,Z,2S	0.50	2Z,S,2Z,+,-	1.50	2Z,S,2Z,a,-	0.90
	2Z,2S,-,Z,S	0.40	6Z,S	0.20	2Z,+,-,Z,2S	1.40	3Z,S,Z,2S	0.73
	3Z,2S,-,S	0.20	2Z,S,Z,3S	0.20	4Z,S,Z,S	0.90	3Z,2S,Z,S	0.57
	2Z,S,Z,3S	0.20	Z,2S,Z,2S,-	0.10	3Z,S,Z,2S	0.80	2Z,S,2Z,+,-	0.50
	2Z,3S,Z,S	0.20	5Z,2S	0.10	2Z,a,2-,2S	0.70	2Z,+,-,Z,2S	0.47
	2Z,2S,-,2S	0.20	3Z,+,2Z,S	0.10	3Z,2S,Z,S	0.60	6Z,S	0.23
	Z,S,Z,S,-,2S	0.10	3Z,2S,-,S	0.10	3Z,+,-,Z,S	0.50	2Z,a,2-,2S	0.23
	6Z,S	0.10	2Z,S,Z,S,-,S	0.10	2Z,2S,2Z,+	0.50	3Z,+,-,Z,S	0.17
	2Z,S,Z,-,2S	0.10	2Z,3S,Z,S	0.10	6Z,S	0.40	2Z,S,Z,3S	0.17
TOTAL (%)	16/17	99.90	16/16	100.00	16/22	99.30	16/31	99.00

Figure 6.15: The most important conformations from CD7 "in vacuo" MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.4.2 Overlapping Analysis I: Accumulative Lambda Index

The *accumulative lambda index* [Fig. 6.16b] is calculated from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 6.16a].

- All the three trajectories; t1, t2 and t3, evolve in time towards better explorations of the conformational space, although –apparently- none of them seems to saturate [Fig. 6.16a]. The individual and combined trajectories suggest that the system is close to a dynamic equilibrium and this point is also confirmed by the saturation plot of lambda index vs. time.
- The total number of conformations in the combined trajectory and the lambda index saturate faster than the individual trajectories, so they can be regarded as good control parameters for monitoring the evolution of the system in time.

- The three trajectories t1, t2 and t3 cannot separately explain the whole system; however, the combined information is still helpful in describing the conformational space.
- Trajectories t1, t2 and t3 have in common about 65% of their conformational spaces, which is nearly 2 parts in 3.

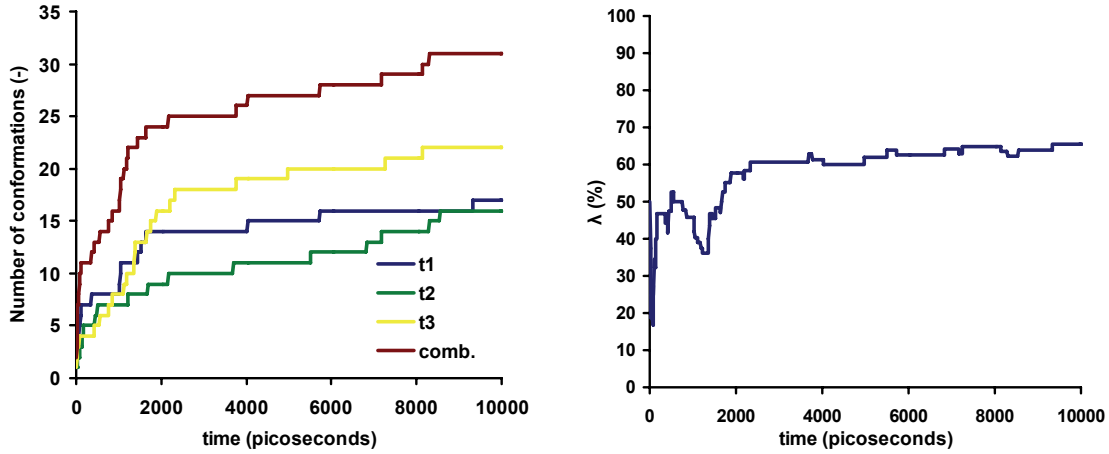


Figure 6.16: (a) Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD7.

Generally speaking, it can be seen that the individual exploration of the conformational space is better in the order t2 < t1 < t3. However, this behaviour is different at the beginning –where t3 starts being the worst although rapidly grows better- and at the end –where t1 saturates and t2 almost crosses the line-.

Detailed information regarding particular conformations found and not found in each trajectory is omitted here because of the considerable amount of conformations involved. Nevertheless, examining the table in [Fig. 6.15], we find again the double behaviour already explained during the simulated annealing calculations:

- On the one hand, a few highly populated conformations are found in t1, t2, t3 and the combined trajectory; [**2Z,S,Z,S,Z,S**], [**2Z,2S,2Z,S**] and [**3Z,S,2Z,S**]: The *Main Conformations*.
- On the other hand, the number of lowly rated non-representative conformations –those below 10%- has increased: The *Conformational Noise*.

This result is confirmed checking the omega index table [Fig. 6.17] where highly rated conformations –those found in t1, t2 and t3- overlap and have high omega indexes. And poorly rated conformations –those that do not overlap in different trajectories- have omega values of 0% in some trajectories.

6.2.4.3 Overlapping Analysis II: Omega Index

The omega indexes have been calculated for the main conformations of CD7 and for some other ones included in the group of conformational noise:

- The three most important conformations of CD7: [**2Z,S,Z,S,Z,S**] with 46.90%, [**2Z,2S,2Z,S**] with 21.17% and [**3Z,S,2Z,S**] with 17.47% experiment overlapping in a rate between 80% and 90%.
- Many other non important conformations –statistically speaking- also overlap. However, the conformations with omega values of 0% are those that have been detected in only a single trajectory but not in the other ones, all at the same time.

name D-II	t1	t2	t3	$<{}^{i}p_j>$	${}^{i}p_{j(max)}$	${}^{i}p_{j(max)}/r$	${}^{i}\nu_j$	${}^{i}d_j$	${}^{i}\omega_j(\%)$
2Z,S,Z,S,Z,S	48.70	51.20	40.80	**46.90**	51.20	17.07	29.83	34.13	**87.40**
2Z,2S,2Z,S	19.90	22.90	20.70	**21.17**	22.90	7.63	13.53	15.27	**88.65**
3Z,S,2Z,S	19.50	15.90	17.00	**17.47**	19.50	6.50	10.97	13.00	**84.36**
Z,2S,Z,S,Z,S	5.10	3.70	2.70	**3.83**	5.10	1.70	2.13	3.40	**62.75**
2Z,2S,Z,2S	2.60	2.40	2.20	**2.40**	2.60	0.87	1.53	1.73	**88.46**
2Z,a,-,Z,2S	0.00	0.00	5.90	**1.97**	5.90	1.97	0.00	3.93	**0.00**
4Z,S,Z,S	1.20	1.80	0.90	**1.30**	1.80	0.60	0.70	1.20	**58.33**
2Z,S,2Z,a,-	0.00	0.00	2.70	**0.90**	2.70	0.90	0.00	1.80	**0.00**
3Z,S,Z,2S	0.90	0.50	0.80	**0.73**	0.90	0.30	0.43	0.60	**72.22**
3Z,2S,Z,S	0.50	0.60	0.60	**0.57**	0.60	0.20	0.37	0.40	**91.67**
2Z,S,2Z,+,-	0.00	0.00	1.50	**0.50**	1.50	0.50	0.00	1.00	**0.00**
2Z,+,-,Z,2S	0.00	0.00	1.40	**0.47**	1.40	0.47	0.00	0.93	**0.00**
6Z,S	0.10	0.20	0.40	**0.23**	0.40	0.13	0.10	0.27	**37.50**
2Z,a,2-,2S	0.00	0.00	0.70	**0.23**	0.70	0.23	0.00	0.47	**0.00**
3Z,+,-,Z,S	0.00	0.00	0.50	**0.17**	0.50	0.17	0.00	0.33	**0.00**
2Z,S,Z,3S	0.20	0.20	0.10	**0.17**	0.20	0.07	0.10	0.13	**75.00**
TOTAL (%)				**99.00**					

Figure 6.17: Omega index –individual conformations overlapping ratio- for CD7.

6.2.4.4 Average Structures

The three *average structures* for CD7 are shown hereafter:

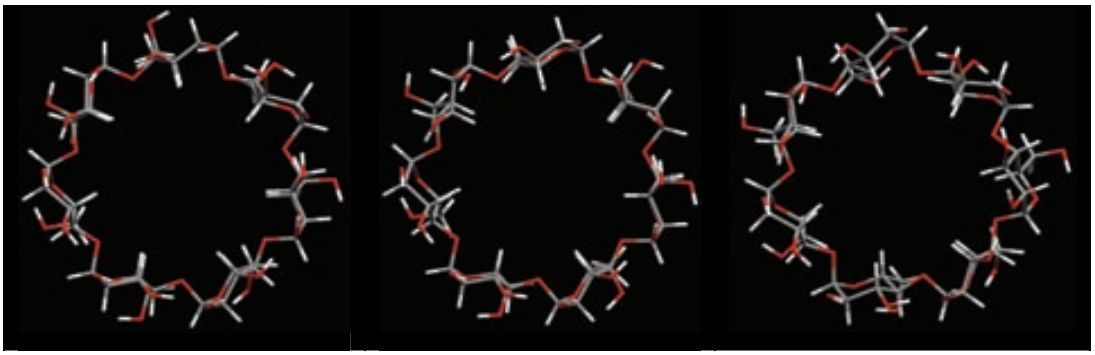


Figure 6.18: Average conformations from CD7 "in vacuo" MD trajectories t1, t2 and t3. The point of view of the images is the face bearing the secondary hydroxyls.

The images in [Fig. 6.18] support the results obtained from the descriptor analysis:

- The average structures from t1, t2 and t3 are viable molecules.
- The remaining rigidity –despite the higher flexibility of CD7- is important in these macrocycles, keeping a cavity available for host-guest phenomena. Solvent Accessible Surface Area calculations show that the cavity is bigger than that of CD6 and encapsulating properties have been widely described in the scientific journals as explained in chapter III.
- Regarding hydrogen bonds; on the one hand, the three average structures from t1, t2 and t3 arrange their primary hydroxyls in a clockwise chain of hydrogen bonds (although one of them seem to be hidden in t2). On the other hand, the average structures from t1, t2 and t3 arrange their secondary hydroxyls in a counterclockwise chain of hydrogen bonds.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms cannot be detected by D-I or D-II; however, the overall behaviour of the macrocycle is still well described.
- As a general property of the small cyclodextrins, the long chains of hydrogen bonds detected in gas phase calculations involving primary and secondary hydroxyls are important to stabilize the macrocycle and therefore those conformations that allow them are the ones preferred.

6.2.5 CD8 (γ-CD)

The set of Molecular Dynamics carried out for CD8 includes four different starting conformations instead of the three conformations used in CD7 and the two employed in

the CD6 and CD5 cases. Again, the high flexibility in CD8 strongly recommended this update in the protocol. The group of the starting conformations included: the first one [**Z,S,Z,S,Z,S,Z,S**]; the second one [**2Z,a,-,Z,S,Z,S**]; the third one [**2Z,2S,2Z,a,-**] and the fourth one [**3Z,2S,-,2S**]. All of them were selected according to statistical weight criteria.

6.2.5.1 Population Analysis

The most representative results can be found in [Fig. 6.19]. In this case, 14 conformations have been selected and explain, in all cases, over 89% of the system. Some important conclusions derived from the table are shown hereafter:

- A total number of 99 conformations were found.
- The trajectories t1, t2, t3, t4 and the combined trajectory –the average- share in common some of the most important conformations, however, they are not necessarily in the first places, in all the trajectories, at the same time.
- The total number of conformations in t1, t2, t3, t4 and the combined trajectory grows as the number of glucoses in the cyclodextrin is increased. In this sense, t1 and t4 present a similar number of conformations –in fact 38 and 33 respectively- followed by t2 –which is slightly bigger with 43 ones- and the most populated; t3 –with 64 ones- being the total number of conformations in the combined trajectory –a total of 99 ones- considerably bigger in comparison to the individual trajectories.
- The most highly rated conformations in t1, t2, t3, t4 and the combined trajectory are: [**2Z,2S,2Z,2S**], [**2Z,S,2Z,S,Z,S**] and [**Z,S,Z,S,Z,S,Z,S**].
- Some of the most populated conformations detected by Molecular Dynamics in the combined trajectory are coincident with the highly populated ones found in the SA Conformational Search. However, conformations [**Z,S,Z,S,Z,S,Z,S**] –rated 3rd during MD calculation- and [**2Z,2S,Z,S,Z,S**] –rated 4th during MD calculations- were found interchanged –positions 4th and 3rd- in Simulated Annealing, meaning that rankings change for those conformations beyond the group of the highly populated ones.

- The starting conformation in t1 is the second most populated conformation. The starting conformation in t2 is not detected in the whole trajectory. The starting conformation in t3 is rated the 5th in the trajectory population analysis. And the starting conformation in t4 is rated 10th. The variety of possibilities seems to suggest that MD calculations do not necessarily get stuck in the starting conformations, although the partial overlapping suggests that every trajectory could cover areas that preferably surround the starting position.

Trajectory	1		2		3	
CD8	name D-II	w (%)	name D-II	w (%)	name D-II	w (%)
Starting Conf.	Z,S,Z,S,Z,S,Z,S	(-)	2Z,a,-,Z,S,Z,S	(-)	2Z,2S,2Z,a,-	(-)
MD Conf.	2Z,2S,2Z,2S	49.90	2Z,2S,2Z,2S	61.10	2Z,S,2Z,S,Z,S	21.50
	Z,S,Z,S,Z,S,Z,S	23.80	Z,S,Z,S,Z,S,Z,S	8.00	2Z,2S,2Z,2S	19.30
	2Z,2S,Z,S,Z,S	4.30	2Z,2S,Z,S,Z,S	6.50	3Z,S,2Z,2S	8.90
	2Z,S,Z,S,Z,2S	2.90	2Z,S,Z,S,Z,2S	3.70	3Z,2S,2Z,a	6.90
	Z,S,Z,S,Z,S,-,S	2.50	2Z,2S,-,S,Z,S	2.80	2Z,2S,2Z,a,-	6.30
	Z,2S,-,2S,Z,S	1.90	Z,2S,-,2S,Z,S	2.00	Z,S,Z,S,Z,S,Z,S	5.50
	2Z,2S,-,S,Z,S	1.90	2Z,2S,Z,-,2S	1.90	2Z,2S,Z,S,Z,S	5.50
	2Z,2S,Z,-,2S	1.70	2Z,S,2Z,S,Z,S	1.70	2Z,S,Z,S,Z,2S	5.00
	2Z,3S,-,2S	1.50	Z,2S,-,2S,-,S	1.40	3Z,2S,2Z,S	3.20
	Z,2S,Z,2S,Z,S	1.40	2Z,2S,-,Z,2S	1.10	3Z,S,Z,S,Z,S	2.20
	2Z,S,2Z,S,Z,S	1.40	Z,2S,Z,2S,Z,S	1.00	3Z,S,3Z,S	1.80
	2Z,S,Z,S,-,2S	1.00	2Z,S,Z,S,-,2S	0.90	Z,2S,-,Z,2S,-	1.20
	Z,2S,-,2S,-,S	0.80	3Z,S,2Z,2S	0.80	Z,2S,-,S,Z,S,-	1.10
	3Z,S,2Z,2S	0.70	3Z,2S,Z,-,+	0.80	2Z,3S,2Z,a	0.90
TOTAL (%)	14/38	95.70	14/43	93.70	14/64	89.30

Trajectory	4		average	
CD8	name D-II	w (%)	name D-II	w (%)
Starting Conf.	3Z,2S,-,2S			
MD Conf.	2Z,2S,2Z,2S	32.80	2Z,2S,2Z,2S	40.78
	2Z,S,2Z,S,Z,S	27.60	2Z,S,2Z,S,Z,S	13.05
	3Z,S,2Z,2S	8.50	Z,S,Z,S,Z,S,Z,S	10.30
	2Z,2S,Z,S,Z,S	5.00	2Z,2S,Z,S,Z,S	5.33
	3Z,2S,2Z,S	4.50	3Z,S,2Z,2S	4.73
	2Z,S,Z,S,Z,2S	4.40	2Z,S,Z,S,Z,2S	4.00
	Z,S,Z,S,Z,S,Z,S	3.90	3Z,2S,2Z,S	2.00
	3Z,S,Z,S,Z,S	2.90	3Z,2S,2Z,a	1.73
	3Z,S,3Z,S	1.50	2Z,2S,2Z,a,-	1.65
	3Z,2S,-,2S	1.20	2Z,2S,-,S,Z,S	1.43
	2Z,S,Z,2S,Z,S	1.20	3Z,S,Z,S,Z,S	1.40
	4Z,S,2Z,S	1.00	Z,2S,-,2S,Z,S	0.98
	2Z,2S,-,S,Z,S	0.60	2Z,2S,Z,-,2S	0.90
	Z,S,-,S,Z,S,-,S	0.50	3Z,S,3Z,S	0.83
TOTAL (%)	14/33	95.60	14/99	89.08

Figure 6.19: The most important conformations from CD8 "in vacuo" MD trajectories 1, 2, 3, 4, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.5.2 Overlapping Analysis I: Accumulative Lambda Index

The *accumulative lambda index* [Fig. 6.20b] is calculated from the total number of conformations in t1, t2, t3, t4 and the combined trajectory [Fig. 6.20a].

- All four trajectories; t1, t2, t3 and t4, evolve in time towards better explorations of the conformational space, although –apparently- none of them seems to reach full saturation [Fig. 6.20a]. As in the case of CD7, the individual and combined trajectories suggest that the system is close to a dynamic equilibrium. This point is confirmed by the saturation plot of lambda index vs. time, although the plot of total number of conformations vs. time is not fully converged.
- The total number of conformations in the combined trajectory and the lambda index saturate faster than the individual trajectories, so they can be regarded as good control parameters for monitoring the evolution of the system in time.
- The four individual trajectories t1, t2, t3 and t4 cannot separately explain the whole system; however, the combined information is still helpful in describing the conformational space.
- Trajectories t1, t2, t3 and t4 have in common about 59% of their conformational spaces, which is nearly 3 parts in 5.

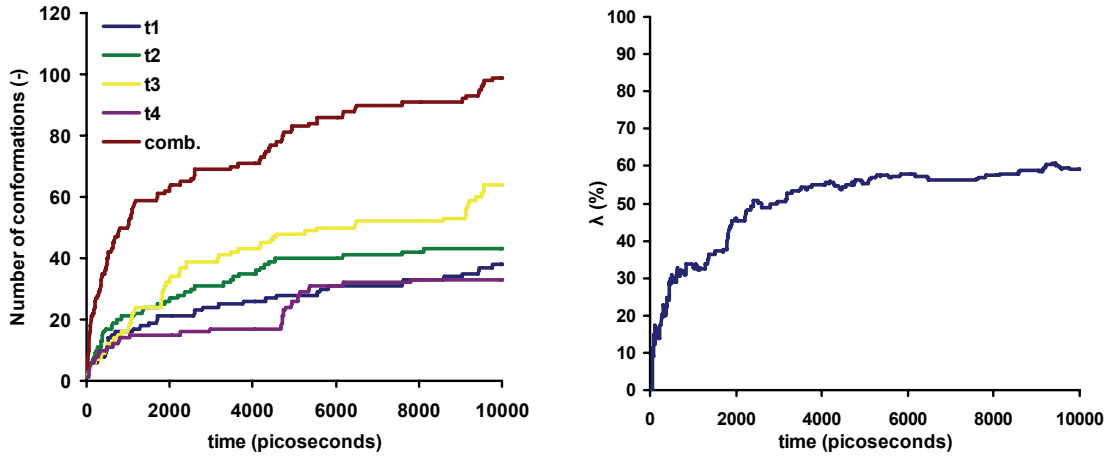


Figure 6.20: (a) Number of conformations in *t1, t2, t3, t4* and total number of conformations in the combined trajectory, *comb.* (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD8.

The efficiency in the individual exploration of conformational space follows the order t4 ≈ t1 < t2 < t3. The main remarkable phenomena in the series are; on the one hand, the spectacular increment in the number of conformations detected in t4 between 4900 and 5200 picoseconds; and on the other hand, a similar effect in t1 detected over 9400 picoseconds. Both changes suggest that the individual trajectories probably discovered new areas in the conformational space not previously visited.

Detailed information regarding particular conformations is omitted here because of the considerable amount of conformations involved in the ensembles. Nevertheless, examining the table in [Fig. 6.19], we find again the double behaviour already explained during the Simulated Annealing calculations:

- On the one hand, a few highly populated conformations are found in t1, t2, t3, t4 and the combined trajectory; [**2Z,2S,2Z,2S**], [**2Z,S,2Z,S,Z,S**] and [**Z,S,Z,S,Z,S,Z,S**]: The *Main Conformations*.
- On the other hand, the number of lowly rated non-representative conformations –those below 10%- has increased: The *Conformational Noise*.

6.2.5.3 Overlapping Analysis II: Omega Index

The previous result is confirmed by checking the omega index table [Fig. 6.21] where highly rated conformations –the main conformations found in t1, t2, t3 and t4- overlap and have high omega indexes, both at the same time. Poorly rated conformations –those that do not overlap in different trajectories- have omega values of 0% in some trajectories. However, the last behaviour is partially biased since the omega index table has been calculated for the most representative conformations of CD8 and only conformation number 8 is rated 0%. Most conformations exhibiting omega indexes of 0% -the conformational noise- were so poorly weighted that they were discarded in the table:

- The three most important conformations of CD8: [**2Z,2S,2Z,2S**] with 40.78%, [**2Z,S,2Z,S,Z,S**] with 13.05% and [**Z,S,Z,S,Z,S,Z,S**] with 10.30% experiment overlapping in rates between 24% and 56%.
- Many other non important conformations –statistically speaking- also overlap. However, the conformations with omega values of 0% are those that have been detected in only a single trajectory but not in the other ones, all at the same time.

name D-II	t1	t2	t3	t4	$<{}^{i}p_{j}>$	${}^{i}p_{j(max)}$	${}^{i}p_{j(max)}/r$	${}^{i}v_{j}$	${}^{i}d_{j}$	${}^{i}\omega_{j}(\%)$
2Z,2S,2Z,2S	49.90	61.10	19.30	32.80	**40.78**	61.10	15.28	25.50	45.83	**55.65**
2Z,S,2Z,S,Z,S	1.40	1.70	21.50	27.60	**13.05**	27.60	6.90	6.15	20.70	**29.71**
Z,S,Z,S,Z,S,Z,S	23.80	8.00	5.50	3.90	**10.30**	23.80	5.95	4.35	17.85	**24.37**
2Z,2S,Z,S,Z,S	4.30	6.50	5.50	5.00	**5.33**	6.50	1.63	3.70	4.88	**75.90**
3Z,S,2Z,2S	0.70	0.80	8.90	8.50	**4.73**	8.90	2.23	2.50	6.68	**37.45**
2Z,S,Z,S,Z,2S	2.90	3.70	5.00	4.40	**4.00**	5.00	1.25	2.75	3.75	**73.33**
3Z,2S,2Z,S	0.20	0.10	3.20	4.50	**2.00**	4.50	1.13	0.88	3.38	**25.93**
3Z,2S,2Z,a	0.00	0.00	6.90	0.00	**1.73**	6.90	1.73	0.00	5.18	**0.00**
2Z,2S,2Z,a,-	0.00	0.30	6.30	0.00	**1.65**	6.30	1.58	0.08	4.73	**1.59**
2Z,2S,-,S,Z,S	1.90	2.80	0.40	0.60	**1.43**	2.80	0.70	0.73	2.10	**34.52**
3Z,S,Z,S,Z,S	0.40	0.10	2.20	2.90	**1.40**	2.90	0.73	0.68	2.18	**31.03**
Z,2S,-,2S,Z,S	1.90	2.00	0.00	0.00	**0.98**	2.00	0.50	0.48	1.50	**31.67**
TOTAL (%)					**87.35**					

Figure 6.21: Omega index –individual conformations overlapping ratio- for CD8.

6.2.5.4 Average Structures

The four *average structures*, those that better describe the overall molecular behaviour in time, are shown hereafter:

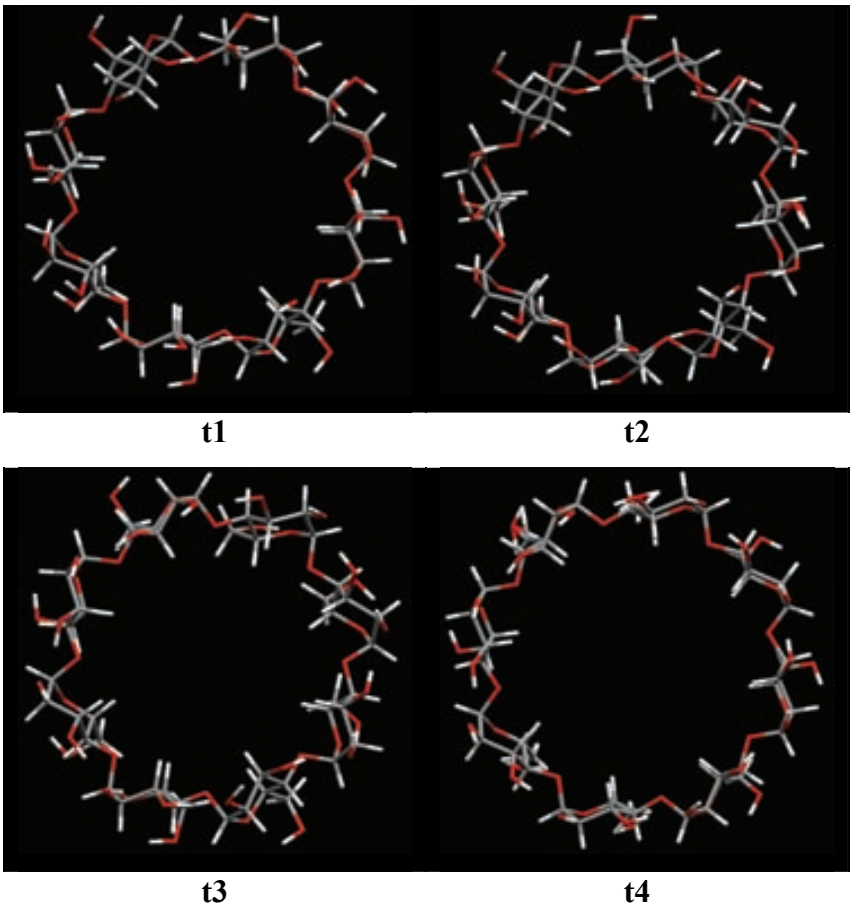


Figure 6.22: Average conformations from CD8 “in vacuo” MD trajectories t1, t2, t3 and t4. The point of view of the images is the face bearing the secondary hydroxyls.

The images [6.22] support the results obtained from the descriptor analysis:

- The average structures from t1, t2, t3 and t4 are viable molecules.
- It can be seen that the cavity is clearly maintained, a consequence of the remaining rigidity; however, flexibility in CD8 is –as the lambda index states- the highest in the group of the small cyclodextrins. The cavity is therefore available for host-guest phenomena already described in chapter III. In this sense, Solvent Accessible Surface Area calculations show that the cavity is bigger than that of CD7.
- The four average structures from t1, t2, t3 and t3 arrange their primary hydroxyls in a clockwise chain of hydrogen bonds (although some of them seem to be partially hidden). The situation is different when describing the secondary hydroxyls: on the one hand, the average structures from t1 and t2 arrange them clockwise; on the other hand, t3 and t4 arrange their secondary hydroxyls in a counterclockwise chain of hydrogen bonds.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms cannot be detected by D-I or D-II; however, the overall behaviour of the macrocycle is still well described.
- As a general property of the small cyclodextrins, the long chains of hydrogen bonds detected in gas phase calculations involving primary and secondary hydroxyls are important to stabilize the macrocycle and therefore those conformations that allow them are the ones preferred.

6.2.6 CD14

This molecule is one of the four cyclodextrins –alongside with CD21, CD26 and CD28- that belongs to the group of the *Large Cyclodextrins*, meaning that the extreme flexibility –already detected and mentioned during the Simulated Annealing calculations- is a determinant factor in their conformational behaviour.

The series of SA conformational search calculations carried out on CD14, CD21, CD26 and CD28 proved that there were no preferred conformations in the ensembles but, mainly, *molecular noise*. Then, it was decided to randomly define a unique number of

conformations for all of them –the final number was 3 conformations- and analyse the behaviour of the three trajectories.

Therefore, the set of Molecular Dynamics carried out for CD14 included three different starting conformations instead of the four conformations used in CD8: the first one [**2Z,S,Z,S,Z,2S,2Z,S,-,2S**]; the second one [**3Z,S,Z,S,Z,S,Z,S,Z,S,Z,S**] and the third one [**3Z,-,S,Z,2S,Z,a,Z,3S**]. All of them were randomly selected.

6.2.6.1 Population Analysis

The most representative results can be found in [Fig. 6.23]. In this case, 8 conformations have been selected. They explain between 38% and 71% of the system, which is, unfortunately, a not so good result. Some conclusions are shown hereafter:

- A total number of 371 conformations were found.
- The total number of conformations found in the individual trajectories t1, t2 and t3 is similar –respectively 104, 157 and 116 conformations- however, the total number of conformations found in the combined trajectory is nearly three times bigger –about 371-.
- The trajectories t1, t2, t3 and the combined trajectory –the average- have in common almost none of the conformations, however, a small amount of them slightly overlap.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**Z,2S,-,Z,S,Z,S,-,2S,Z,+,-**], [**2Z,S,2Z,S,-,2S,Z,S,Z,2S**], [**Z,S,Z,S,-,2S,Z,+,-,3S,-**] and [**2Z,2S,-,Z,2S,Z,2S,2Z,S**]. Nevertheless, their ratios are lower than those measured for the small cyclodextrins.
- There is no correspondence between the most populated conformations detected by Molecular Dynamics in the combined trajectory and the highly populated ones in the SA Conformational Search.
- The starting conformations in t1, t2 and t3 are not detected during the MD calculations; however, there are highly rated conformations in every individual trajectory. The result suggests that, on the one hand, MD calculations do not

necessarily get stuck in the starting conformations, and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1		2	
CD14	name D-II	w (%)	name D-II	w (%)
Starting Conf.	2Z,S,Z,S,Z,2S,2Z,S,-,2S	(-)	3Z,S,Z,S,Z,S,Z,S,Z,S,Z,S	(-)
MD Conf.	2Z,S,2Z,S,-,2S,Z,S,Z,2S	22.50	2Z,2S,-,Z,2S,Z,2S,2Z,S	12.20
	2Z,S,Z,S,-,Z,4S,Z,2S	11.10	2Z,S,-,2S,Z,3S,Z,S,Z,S	6.70
	2Z,4S,Z,2S,2Z,S,Z,S	11.10	Z,2S,Z,2S,Z,S,Z,S,Z,2S,-	5.90
	2Z,4S,Z,2S,-,Z,S,Z,S	5.90	Z,3S,Z,S,Z,S,Z,S,Z,-,2S	5.80
	Z,4S,Z,2S,-,Z,S,Z,S,-	4.90	2Z,S,Z,S,-,2S,-,2S,Z,2S	5.50
	2Z,2S,Z,S,Z,2S,2Z,S,Z,S	3.40	2Z,S,-,3S,-,Z,2S,Z,2S	4.20
	2Z,S,Z,2S,-,2S,Z,S,Z,2S	3.30	Z,3S,Z,S,Z,S,-,2S,-,Z,S	3.10
	2Z,4S,Z,2S,2Z,S,-,S	2.30	Z,2S,Z,2S,Z,S,-,S,Z,2S,-	3.10
TOTAL (%)	8/104	64.50	8/157	46.50

Trajectory	3		average	
CD14	name D-II	w (%)	name D-II	w (%)
Starting Conf.	3Z,-,S,Z,2S,Z,a,Z,3S	(-)		
MD Conf.	Z,2S,-,Z,S,Z,S,-,2S,Z,+,-	26.80	Z,2S,-,Z,S,Z,S,-,2S,Z,+,-	8.93
	Z,S,Z,S,-,2S,Z,+,-,3S,-	13.00	2Z,S,2Z,S,-,2S,Z,S,Z,2S	7.50
	2Z,S,Z,S,-,2S,Z,+,-,Z,2S	10.60	Z,S,Z,S,-,2S,Z,+,-,3S,-	4.33
	2Z,S,Z,S,-,2S,Z,+,-,3S	6.10	2Z,2S,-,Z,2S,Z,2S,2Z,S	4.07
	2Z,S,Z,S,-,Z,S,Z,+,-,3S	5.60	2Z,S,Z,S,-,Z,4S,Z,2S	3.70
	2Z,+,-,Z,2S,-,Z,S,Z,S,-,S	4.40	2Z,4S,Z,2S,2Z,S,Z,S	3.70
	Z,S,Z,S,-,Z,S,Z,+,-,3S,-	2.50	2Z,S,Z,S,-,2S,Z,+,-,Z,2S	3.53
	2Z,+,-,3S,-,Z,S,Z,S,-,S	2.10	2Z,S,-,2S,Z,3S,Z,S,Z,S	2.23
TOTAL (%)	8/116	71.10	8/371	38.00

Figure 6.23: The most important conformations from CD14 "in vacuo" MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.6.2 Overlapping Analysis I: Accumulative Lambda Index

The *accumulative lambda index* [Fig. 6.24b] is calculated from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 6.24a].

- All the three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them reaches full saturation [Fig. 6.24a]. In this case, the saturation plot of lambda index vs. time [Fig. 6.24b] – scaled to 10%- is not trustworthy: firstly, it is not sure that it fully converges; secondly, the ratios are really low; and finally, the tendency is errant.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since

none of them saturates, information about the conformational space is unreliable.

- Trajectories t1, t2, and t3 have in common –at most- about 3% of their conformational spaces, which is nearly 1 part in 33. On the grounds of these results, the system is arguably underexplored.

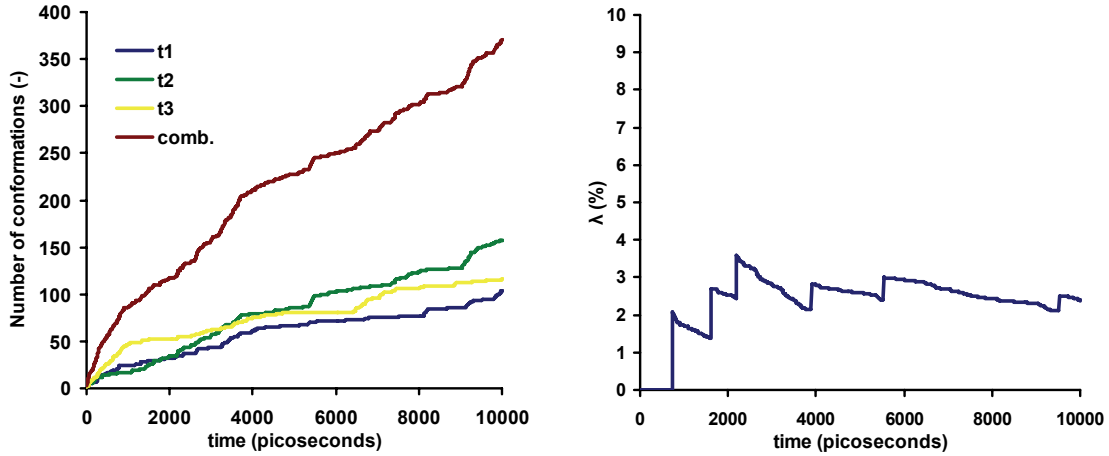


Figure 6.24: (a) Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD14.

The three trajectories show similar efficiencies in the individual exploration of conformational space, although the tendency t2 > t3 > t1 is detected. The most remarkable phenomenon in the series is that the combined trajectory has almost three times the number of conformations found in t1, t2 or t3. This means that every individual trajectory covered separated areas in conformational space, which explains the extremely low lambda index ratio.

Examining the table in [Fig. 6.23], it is found that the double behaviour previously detected in CD7 and CD8 gas phase MD and SA calculations –the main conformations and the conformational noise- is not so clearly observed. However, it is still there, though in a wider sense. Focusing the conformational behaviour under the optics of

Pareto's Law –or the 80/20 Law[197]- [Fig. 6.25] the trajectories could still be divided into the two groups:

CD14		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	104	157	116	371
20% of Conf.	(n)	21	31	23	74
Acc. Stat. Weight	(%)	83.70	76.10	83.30	81.20

Figure 6.25: Pareto. CD14 MD trajectories in gas phase.

The best 20% conformations represent between 76% and 84% of their trajectories which agrees with the results explained in [Fig. 6.23]: the first 8 conformations of t1, t2, t3, and the combined trajectories were shown with their statistical weights: 64%, 46%, 71% and 38%. These ratios suggest that they can be considered as the *main conformations* –in the individual trajectories- while the others are the *conformational noise*.

6.2.6.3 Overlapping Analysis II: Omega Index

The omega index was 0% for almost all the conformations found, so the table was omitted. The extremely low overlapping ratio between trajectories –the lambda index value is only about 3%- suggests that only a few poorly rated conformations, located at the outer parts of the orbits, *touched* other trajectories.

6.2.6.4 Average Structures

The group of the large cyclodextrins is not only quantitatively but also qualitatively different from the group of the small cyclodextrins: as was already said, the larger ones can fold while the smaller ones cannot.

This change in conformational behaviour was so important that it suggested us to re-scale our point of view towards protein-like macromolecules and approach the conformational study of cyclodextrins under the optics of a new criterion in a more general way: trying new visualizations like strands, ribbons and bands; less precise than descriptors D-I and D-II, although more general.

[197] ***Pareto's Law*** is a quasi-empirical statement –commonly used in social sciences, economics and politics- that read: "In any system, 20% of the individuals generally explains 80% of the properties, while the other 80% of individuals explains the remaining 20% of the properties".

Therefore, the images combining molecules with ribbons in [Fig. 6.26] and other ones in further sections describing *large cyclodextrins* have been created with ***gOpenMol molecular viewer***[198]. A PDB file containing the average structure was previously computed with AMBER 7 module ptraj processing the MD trajectory file; t1 and t2. The command file can be consulted in the DVD: appendix 10.3.1.2 (chapter X Amber 7 files). The ribbon was selected to sequentially link the oxygen atoms responsible for the glycosidic bonds.

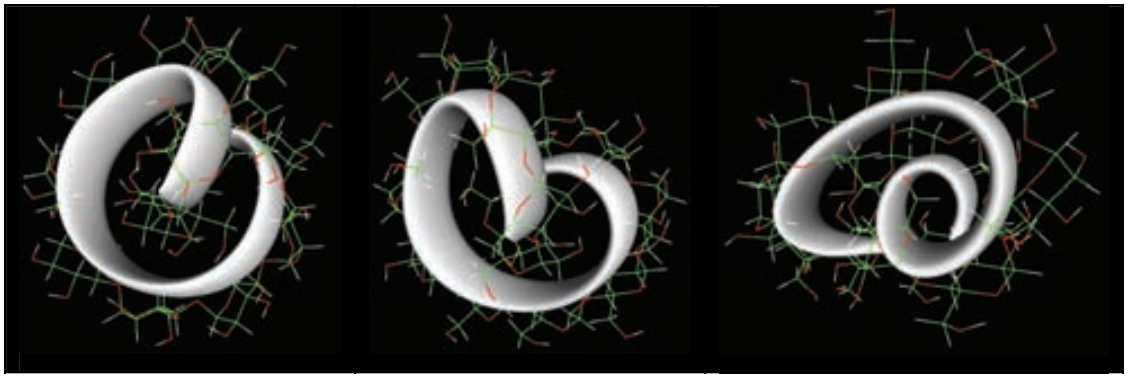


Figure 6.26: Average conformations from CD14 "in vacuo" MD trajectories t1, t2 and t3.

The images of the average structures [Fig. 6.26] seem to contradict the results obtained from the descriptor analysis: On the one hand, the analysis of D-II –lambda index, omega index, individual and total number of conformations- leads to CD14 as a highly flexible molecule with a wide conformational space. On the other hand, and according to the new criterion –the one based on the visual comparison of the conformations in the ribbon representation-, the three average structures suggest that the three molecules are mostly the same –although the point of view in t3 has been changed to show the loop-.

The opposing results reveal that D-II is not probably the best descriptor for large cyclodextrins when folding is detected; D-II is unable to effectively model large molecules that fold in space. The difference between them is really important and the scale factor must not be regarded as a secondary topic. Under this new optics, the cyclodextrins in the present work have been divided just in two families: those that can fold, and those that cannot.

[198] **gOpenMol – A Molecular Viewer** (a) Laaksonen, L.; *J. Mol. Graph*. **1992**. *10(1)*. 33-34. (b) Bergman, D.L.; Laaksonen, L.; Laaksonen, A.; *J. Mol. Graph. Model*. **1997**. *15(5)*. 301-306.
(c) http://www.csc.fi/english/pages/g0penMol

These results –as was said- strongly recommend developing a new descriptor –more general and less precise- that should specifically take into account folding in space and correctly describe the secondary structure.

Now, specifically addressing to CD14 average structures:

- The average structures from t1, t2 and t3 –similarly to the case of the small cyclodextrins- are viable molecules.
- Under the new criterion, all three of them, t1, t2 and t3, can be regarded as similar conformations. In all three cases, CD14 folds creating a small internal loop stabilised by intramolecular interactions.
- Hydrogen bonds are also very important to stabilise conformations, however, the long chains separately involving primary and secondary hydroxyls –typical of the small CDs- now coexist with those hydrogen bonds involving distant glucoses in the primary structure, now close in space because of the secondary structure.
- The directionality –clockwise and counterclockwise- involving hydrogen bonds between hydroxyls and oxygen atoms is a less important phenomenon not fully used in the family of the large CDs. Partial sequences involving certain number of glucoses are obviously detected; nevertheless, the length of the chain rarely includes the total length of the macroring.
- Hydrogen bonds –considered in general- involving primary and secondary hydroxyls are important to stabilize a large number of folded conformations in space.

6.2.7 CD21

This molecule is also one of the four cyclodextrins that belong to the group of the *Large Cyclodextrins*, meaning that the extreme flexibility is a determinant factor in their conformational behaviour. As in the case of CD14, three conformations were selected for the MD calculations. Therefore, the set of Molecular Dynamics carried out for CD21 included three different starting conformations: the first one [**5Z,S,2Z,2S,Z,-**

,Z,S,Z,S,Z,S,Z,2S]; the second one [**2Z,a,-,Z,S,-,S,Z,S,Z,a,Z,3S,-,Z,S,-,S**] and the third one [**2Z,-,S,Z,2S,-,S,Z,S,Z,2S,a,2S,-,a,-,S**]. All of them were randomly selected.

6.2.7.1 Population Analysis

The most representative results can be found in [Fig. 6.27]. In this case, as was done in the case of CD14, 8 conformations have been selected. They explain between 21% and 46% of the system, which not only is, unfortunately, a not very good result but also a situation worse than that shown in CD14.

- A total number of 763 conformations were found.
- The total number of conformations found in the individual trajectories t1 and t3 is similar –204 and 209 respectively-. Trajectory t2 is, however, noticeably wider with 350 being almost twice t1 or t3. Anyway, the total number of conformations found in the combined trajectory is nearly three times bigger –about 763-.
- The trajectories t1, t2, t3 and the combined trajectory have no common conformations.
- The ratios of the *most rated* conformations decrease as the number of glucoses grows.
- The most highly rated conformations in t1, t2 and t3 are: [**3Z,S,-,Z,S,Z,3S,Z,-,Z,S,-,S,2Z,2S**], [**Z,4S,-,Z,3S,-,a,S,Z,S,Z,S,-,S,-,S**] and [**2Z,2S,a,S,2Z,a,-,2Z,-,Z,2S,Z,S,-,2S**], and they are not necessarily the same as the highly rated ones in the combined trajectory since the overlapping ratio is 0%. Furthermore, their ratios have diminished in comparison to the smaller cyclodextrins.
- There is no correspondence at all between the most populated conformations detected by Molecular Dynamics in the combined trajectory and the highly populated ones in the SA Conformational Search.
- Similarly to the case of CD14, the starting conformations in t1, t2 and t3 are not detected during the MD calculations; however, there are highly rated conformations in every individual trajectory. The result suggests that, on the one hand, MD calculations do not necessarily get stuck in the starting conformations,

and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1		2	
CD21	name D-II	w (%)	name D-II	w (%)
Starting Conf.	5Z,S,2Z,2S,Z,-,Z,S,Z,S,Z,S,Z,2S	(-)	2Z,a,-,Z,S,-,S,Z,S,Z,a,Z,3S,-,Z,S,-,S	(-)
MD Conf.	3Z,S,-,Z,S,Z,3S,Z,-,Z,S,-,S,2Z,2S	11.40	Z,4S,-,Z,3S,-,a,S,Z,S,Z,S,-,S,-,S	6.10
	3Z,S,-,Z,S,2Z,2S,Z,-,Z,S,-,S,2Z,2S	9.20	2Z,2S,-,S,-,S,Z,4S,-,Z,3S,-,a,S	5.10
	2Z,2S,Z,-,Z,S,-,Z,S,2Z,2S,Z,-,Z,S,-,S	7.90	2Z,S,Z,S,-,S,Z,+,Z,S,Z,S,2Z,-,+,2S,-,a	3.30
	3Z,S,-,Z,S,Z,3S,3Z,S,-,S,2Z,2S	4.70	2Z,S,Z,S,-,S,Z,a,Z,S,Z,S,2Z,-,+,2S,-,a	3.10
	3Z,S,-,Z,S,-,3S,Z,-,Z,S,-,S,2Z,2S	3.50	2Z,S,Z,S,-,S,Z,a,-,S,Z,S,2Z,-,+,2S,-,a	2.20
	3Z,S,-,Z,S,Z,3S,Z,-,Z,S,-,S,Z,3S	3.40	2Z,S,-,S,-,S,Z,+,Z,S,Z,S,2Z,-,+,2S,-,a	2.10
	3Z,S,-,Z,S,2Z,2S,3Z,S,-,S,2Z,2S	2.90	2Z,-,+,2S,-,a,S,Z,S,Z,S,-,S,-,+,Z,3S	2.00
	3Z,S,-,Z,S,2Z,2S,Z,-,Z,S,-,S,Z,3S	2.80	2Z,2S,-,S,Z,S,Z,4S,-,Z,3S,-,a,S	1.90
TOTAL (%)	8/204	45.80	8/350	25.80

Trajectory	3		average	
CD21	name D-II	w (%)	name D-II	w (%)
Starting Conf.	2Z,-,S,Z,2S,-,S,Z,S,Z,2S,a,2S,-,a,-,S	(-)		
MD Conf.	2Z,2S,a,S,2Z,a,-,2Z,-,Z,2S,Z,S,-,2S	12.90	2Z,2S,a,S,2Z,a,-,2Z,-,Z,2S,Z,S,-,2S	4.30
	2Z,2S,a,S,2Z,a,-,2Z,2-,2S,Z,S,-,2S	6.20	3Z,S,-,Z,S,Z,3S,Z,-,Z,S,-,S,2Z,2S	3.80
	2Z,-,Z,S,Z,2S,-,S,Z,S,Z,2S,a,S,Z,-,a,-	4.00	3Z,S,-,Z,S,2Z,2S,Z,-,Z,S,-,S,2Z,2S	3.07
	2Z,S,Z,a,S,2Z,a,-,2Z,2-,2S,Z,S,Z,2S	3.70	2Z,2S,Z,-,Z,S,-,Z,S,2Z,2S,Z,-,Z,S,-,S	2.63
	2Z,S,Z,a,S,2Z,a,-,2Z,2-,2S,Z,S,-,2S	3.10	2Z,2S,a,S,2Z,a,-,2Z,2-,2S,Z,S,-,2S	2.07
	Z,2S,a,S,Z,-,a,-,S,Z,-,Z,S,Z,2S,-,S,Z,S	2.80	Z,4S,-,Z,3S,-,a,S,Z,S,Z,S,-,S,-,S	2.03
	2Z,2S,a,S,Z,-,a,-,2Z,-,Z,2S,Z,S,-,2S	2.70	2Z,2S,-,S,-,S,Z,4S,-,Z,3S,-,a,S	1.70
	2Z,2S,a,S,2Z,a,-,Z,S,2-,2S,Z,S,-,2S	2.40	3Z,S,-,Z,S,Z,3S,3Z,S,-,S,2Z,2S	1.57
TOTAL (%)	8/209	37.80	8/763	21.17

Figure 6.27: The most important conformations from CD21 "in vacuo" MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.7.2 Overlapping Analysis I: Accumulative Lambda Index

In this case, the *lambda index* is measured at the end of the MD calculations [Fig. 6.28] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 6.29]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD21	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	204	763	763	254.33	**0.00**
	2	350				
	3	209				

Figure 6.28: *Lambda index* –trajectories overlapping ratio- for CD21 at the end of the MD calculations.

- All three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 6.29]. As said, the

saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 6.28], therefore, calculations are not trustworthy: on the one hand, it is not sure that trajectories fully converge; on the second hand, the ratios are really low.

- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturates, information about the conformational space is unreliable.
- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is arguably underexplored.

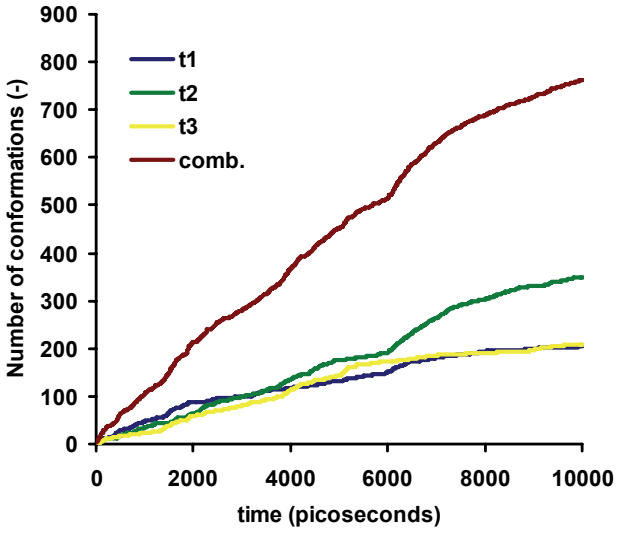


Figure 6.29: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD21.

The three trajectories show a similar efficiency in the individual exploration of conformational space, although the tendency t2 > t3 ≈ t1 is detected. It can be seen that the combined trajectory has nearly three times the number of conformations found in t1, t2 or t3. This means that every individual trajectory covered separated areas in conformational space, which explains the 0% lambda index ratio.

The table in [Fig. 6.27] shows that –step by step, as the number of glucoses grows- the double behaviour is vanishing while the *conformational noise* grows. Anyway, the Pareto diagram –the 80/20 Law- [Fig. 6.30] seems to suggest that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, and decreasing:

CD21		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	204	350	209	763
20% of Conf.	(n)	41	70	42	153
Acc. Stat. Weight	(%)	**77.00**	**65.90**	**72.60**	**72.33**

Figure 6.30: Pareto. CD21 MD trajectories in gas phase.

As said, the best 20% conformations represent between 66% and 77% of their trajectories which agrees with the results explained in [Fig. 6.27] where the first 8 conformations of t1, t2, t3, and the combined trajectories were shown with their statistical weights: 46%, 26%, 38% and 21%. These ratios suggest that they could still be considered as the *main conformations* –in the individual trajectories- while the others could be the *conformational noise*. However, it is important to say that the percentage covered by the best conformations has diminished in comparison to CD14. The noise is ruling the conformational behaviour.

6.2.7.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- suggests that the three trajectories are completely unconnected.

6.2.7.4 Average Structures

CD21 is also one of the large cyclodextrins and then, protein-like visualizations of the type strands, ribbons and bands are employed:

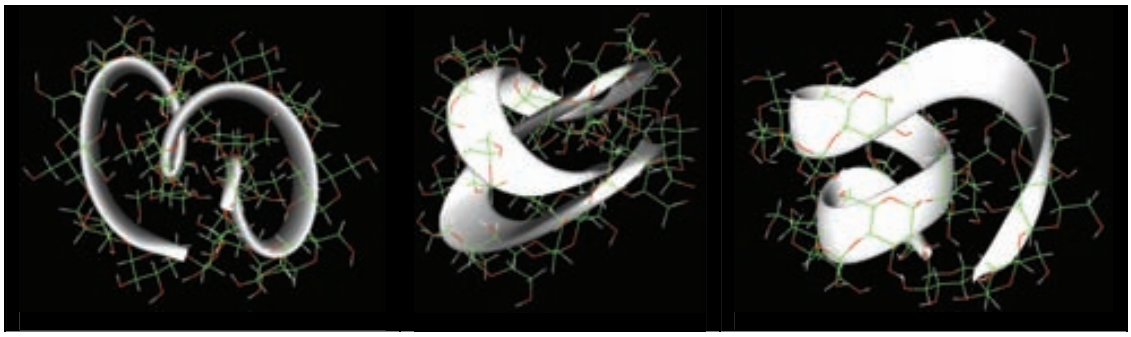


Figure 6.31: Average conformations from CD21 “in vacuo” MD trajectories t1, t2 and t3.

The images of the average structures [Fig. 6.31] again contradict the results obtained from the descriptor analysis. On the one hand, the analysis of D-II states that all conformations in t1, t2 and t3 are different, therefore, the average structures of such ensembles should render *artefacts* similar to those obtained during the SA calculations. Nevertheless, as can be seen, three different viable conformations are obtained instead, meaning that the individual conformations are not so distant in conformational space than was expected.

These results prove that, for our purposes, D-II is extremely sensitive to small conformational changes and therefore is inadequate to describe CD21. Likewise, Molecular Dynamics trajectories tend to remain reasonably close to certain conformational areas not far from the starting conformations.

- Under the new criterion, the three of them; t1, t2 and t3 are also different conformations. In all three cases, CD21 folds creating double internal loops –in t1 and t2- and a tentative kind of helix –in t3-.
- Hydrogen bonds are very important to stabilise conformations. Different types of arrangements coexist at the same time to allow loops, chains and helices because of the secondary structure.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms is a less important phenomenon not fully used in the family of the large CDs. Partial sequences are detected; however, the length of the chain rarely includes the total length of the macroring.

6.2.8 CD26

CD26 is also one of the four cyclodextrins that belongs to the group of the *Large Cyclodextrins*, meaning that the extreme flexibility is again a determinant factor in their conformational behaviour. As in the case of CD14 and CD21, three conformations were selected for the MD calculations. Therefore, the set of Molecular Dynamics carried out for CD26 included three different starting conformations: the first one [**3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S**]; the second one [**2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S**] and the third one [**3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a**]. All of them were randomly selected.

6.2.8.1 Population Analysis

The most representative results can be found in [Fig. 6.32]. As was previously done in the case of CD14 and CD21; 8 conformations have been selected. They explain between 20% and 37% of the system, which is not a very good rate. Besides, the situation is worse than in CD14 and CD21.

- A total number of 760 conformations were found.
- The total number of conformations found in the individual trajectories t1, t2 and t3 is similar –259, 239 and 262 respectively- the total number of conformations in the combined trajectory being nearly three times bigger –about 760-.
- The trajectories t1, t2, t3 and the combined trajectory do not have any conformation in common.
- The ratios of the *most rated* conformations decrease as the number of glucoses grows.
- The most highly rated conformations in t1, t2 and t3 are: [**2Z,2S,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a**], [**3Z,2S,-,2Z,a,-,S,2Z,S,a,-,Z,2S,-,a,3S,Z,S**] and [**3Z,S,Z,-,3S,-,Z,S,2Z,S,Z,4S,-,a,3Z,a**], and they are not necessarily the same as the highly rated ones in the combined trajectory since the overlapping ratio is 0%. Furthermore, their ratios have diminished in comparison to the smaller cyclodextrins.

- There is no correspondence at all between the most populated conformations detected by Molecular Dynamics in the combined trajectory and the highly populated ones in the SA Conformational Search.
- Similarly to the case of CD14 and CD21, the starting conformations in t1, t2 and t3 are not detected during the MD calculations; however, there are highly rated conformations in every individual trajectory. The result suggests that, although MD calculations do not necessarily get stuck in the starting conformations, every trajectory covers areas that preferably surround certain positions.

Trajectory	1	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S	(-)
MD Conf.	2Z,2S,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a	9.40
	3Z,+,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a	8.40
	2Z,2S,2-,Z,3S,Z,S,2Z,S,Z,S,-,Z,2S,-,2S,-,a	5.20
	2Z,2S,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,Z,S,-,a	4.40
	3Z,S,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a	2.60
	2Z,2S,2-,Z,3S,Z,2S,-,S,Z,S,-,Z,2S,-,2S,-,a	2.60
	3Z,+,2-,Z,3S,Z,S,2Z,S,Z,S,-,Z,2S,-,2S,-,a	2.30
	2Z,2S,2-,Z,3S,Z,S,2Z,S,Z,S,-,Z,2S,-,Z,S,-,a	2.30
TOTAL (%)	8/259	37.20

Trajectory	2	
CD26	name D-II	w (%)
Starting Conf.	2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S	(-)
MD Conf.	3Z,2S,-,2Z,a,-,S,2Z,S,a,-,Z,2S,-,a,3S,Z,S	11.80
	3Z,2S,-,2Z,a,-,S,-,Z,S,a,-,Z,2S,-,a,3S,Z,S	5.80
	3Z,2S,-,2Z,a,-,S,Z,-,S,a,-,Z,2S,-,a,3S,Z,S	4.20
	2Z,S,a,-,Z,2S,-,a,3S,Z,S,2Z,-,2S,-,2Z,a,-,S	3.50
	3Z,2S,-,2Z,a,-,S,-,Z,S,a,-,Z,S,2Z,a,3S,Z,S	3.40
	3Z,2S,-,2Z,a,-,3Z,S,a,-,Z,2S,-,a,3S,Z,S	3.30
	3Z,2S,-,2Z,a,-,S,2Z,S,a,2Z,2S,-,a,3S,Z,S	3.20
	3Z,2S,-,2Z,a,-,S,Z,-,S,a,2Z,2S,-,a,3S,Z,S	2.00
TOTAL (%)	8/239	37.20

Trajectory	3	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a	(-)
MD Conf.	3Z,S,Z,-,3S,-,Z,S,2Z,S,Z,4S,-,a,3Z,a	6.40
	3Z,S,Z,-,3S,2-,S,2Z,S,Z,4S,-,a,3Z,a	6.30
	3Z,S,Z,-,3S,-,Z,S,2Z,S,Z,4S,-,a,Z,-,Z,a	5.40
	3Z,S,Z,-,3S,2-,S,2Z,S,Z,4S,-,a,Z,-,Z,a	3.30
	3Z,S,2Z,3S,-,Z,S,2Z,S,Z,4S,-,a,3Z,a	3.30
	3Z,S,2Z,3S,-,Z,S,2Z,S,Z,4S,-,a,Z,-,Z,a	2.60
	3Z,a,-,2Z,S,Z,-,3S,2-,S,2Z,S,Z,4S,Z,a	2.50
	3Z,S,Z,-,3S,-,Z,S,-,Z,S,Z,4S,-,a,3Z,a	2.20
TOTAL (%)	8/262	32.00

Trajectory	average	
CD26	name D-II	w (%)
Starting Conf.		
MD Conf.	3Z,2S,-,2Z,a,-,S,2Z,S,a,-,Z,2S,-,a,3S,Z,S	3.93
	2Z,2S,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a	3.13
	3Z,+,2-,Z,3S,Z,S,Z,-,S,Z,S,-,Z,2S,-,2S,-,a	2.80
	3Z,S,Z,-,3S,-,Z,S,2Z,S,Z,4S,-,a,3Z,a	2.13
	3Z,S,Z,-,3S,2-,S,2Z,S,Z,4S,-,a,3Z,a	2.10
	3Z,2S,-,2Z,a,-,S,-,Z,S,a,-,Z,2S,-,a,3S,Z,S	1.93
	3Z,S,Z,-,3S,-,Z,S,2Z,S,Z,4S,-,a,Z,-,Z,a	1.80
	2Z,2S,2-,Z,3S,Z,S,2Z,S,Z,S,-,Z,2S,-,2S,-,a	1.73
TOTAL (%)	8/760	19.57

Figure 6.32: The most important conformations from CD26 "in vacuo" MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.8.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 6.33] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 6.34]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD26	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	259	760	760	253.33	**0.00**
	2	239				
	3	262				

Figure 6.33: *Lambda index* –trajectories overlapping ratio- for CD26.

- The three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 6.34]. Besides, the saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 6.33]; therefore, calculations are not trustworthy: on the one hand, it is not sure that trajectories fully converge; on the other hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturates, information about the conformational space is unreliable.

- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is arguably underexplored.

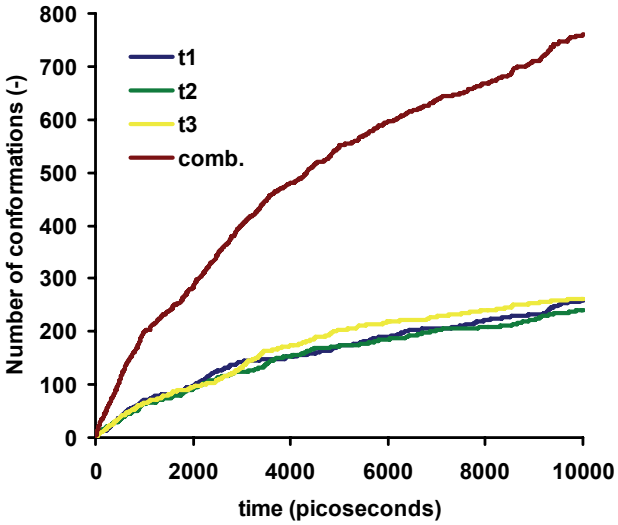


Figure 6.34: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD26.

The three trajectories display a similar efficiency in the individual exploration of conformational space: the general tendency is t1 ≈ t2 ≈ t3. Furthermore, it can be seen that the combined trajectory has nearly three times the number of conformations found in t1, t2 or t3; this means that every individual trajectory covered disconnected areas in conformational space, which, in addition, explains the 0% lambda index ratio.

Similarly to CD14 and CD21, the table in [Fig. 6.32] shows that –as the number of glucoses grows- the two-group behaviour vanishes while the *conformational noise* grows. The Pareto diagram [Fig. 6.35] seems to suggest that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, and similar to the CD21 case:

CD26		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	259	239	262	760
20% of Conf.	(n)	52	48	52	152
Acc. Stat. Weight	(%)	**73.70**	**72.60**	**70.40**	**72.23**

Figure 6.35: Pareto. CD26 MD trajectories in gas phase.

The best 20% conformations represent between 70% and 74% of their trajectories which is in agreement with the results explained in [Fig. 6.32] where the first 8 conformations of t1, t2, t3, and the combined trajectories were shown with their statistical weights: 37%, 37%, 32% and 20%. These ratios –similar to those in CD21- suggest that they could still be considered as the *main conformations* –in the individual trajectories- while the others could be the *conformational noise*. Anyway, as was said in the case of CD21, the noise is ruling the conformational behaviour.

6.2.8.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- suggests that the three trajectories are completely unconnected.

6.2.8.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

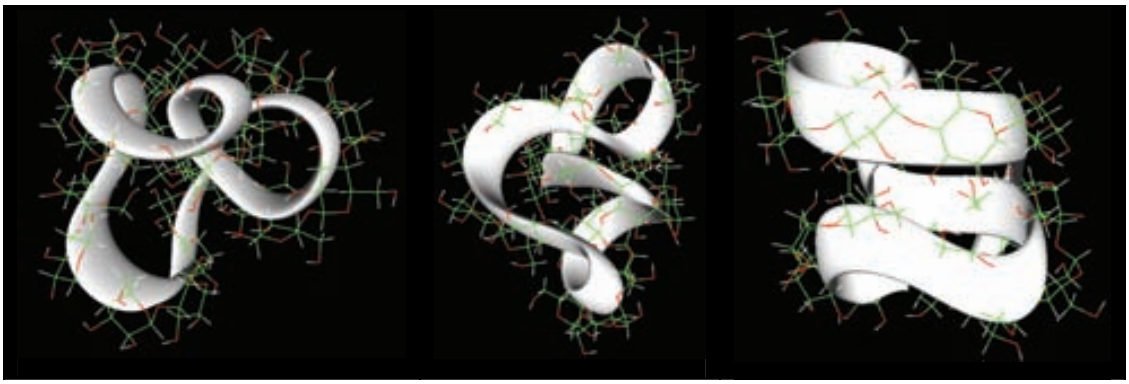


Figure 6.36: Average conformations from CD26 "in vacuo" MD trajectories t1, t2 and t3.

A scenario similar to that of CD21 is met and reasons offered in that section are also applicable here. The images of the average structures [Fig. 6.36] contradict the results obtained from the descriptor analysis suggesting that D-II is, for our purposes, extremely sensitive to small conformational changes and therefore inadequate to

describe CD26. Likewise, Molecular Dynamics trajectories tend to remain reasonably close to the starting conformations.

- Under the new criterion, the three of them; t1, t2 and t3 can be regarded as different conformations. CD26 folds creating a variety of loops and partially closed conformations –in t1 and t2- while a compact helix/ribbon is obtained –in t3-.
- Hydrogen bonds are very important to stabilise conformations. Different types of arrangements coexist at the same time to allow loops, chains and helices because of the secondary structure.
- The directionality involving hydrogen bonds is a less important phenomenon not fully used in the family of the large CDs. However, partial sequences are detected.

6.2.9 CD28

CD28 is the last one –and also the largest- of the four cyclodextrins studied in this work that belongs to the group of the *Large Cyclodextrins*. Similarly to the other cyclodextrins in the group, the extreme flexibility is a determinant factor in its conformational behaviour. As in the case of CD14, CD21 and CD26, three conformations were selected for the MD calculations. Hence, the set of Molecular Dynamics carried out for CD26 included three different starting conformations: the first one [**3Z,S,+,Z,S,Z,a,2Z,S,2Z,2S,a,-,Z,S,-,2Z,S,a,+,a,-**]; the second one [**5Z,a,-,Z,S,Z,S,+,Z,S,-,a,2Z,S,Z,2S,-,a,2S,-,a**] and the third one [**2Z,S,2Z,S,-,a,Z,-,2S,Z,a,2S,-,S,Z,-,S,-,2S,Z,3S**]. All of them were randomly selected.

6.2.9.1 Population Analysis

The most representative results are shown in [Fig. 6.37]. As was previously done in the case of CD14, CD21 and CD26; 8 conformations have been selected. They explain between 27% and 52% of the system, which is not a very good rate, although surprisingly improves the statistical weights found in CD26. Anyway, the situation is worse than in the case of the small cyclodextrins.

- A total number of 721 conformations were found.
- The total number of conformations found in the individual trajectories t1 and t3 is similar –respectively 253 and 256-. Trajectory t2 is, however, slightly narrower with only 212. The total number of conformations in the combined trajectory is nearly three times bigger –about 721-.
- The trajectories t1, t2, t3 and the combined trajectory have no common conformations.
- The ratios of the *most rated* conformations decrease as the number of glucoses grows.
- The most highly rated conformations in t1, t2 and t3 are: [**2Z,S,Z,-,2S,a,-,S,2Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a**], [**5Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a**] and [**3Z,S,-,S,-,a,-,Z,2S,-,Z,2S,3Z,-,Z,2S,Z,S,Z,2S**], and they are not necessarily the same as the highly rated ones in the combined trajectory since the overlapping ratio is 0%. Furthermore, their ratios have diminished in comparison to the smaller cyclodextrins.
- There is no correspondence between the most populated conformations detected by Molecular Dynamics in the combined trajectory and the highly populated ones in the SA Conformational Search.
- Similarly to the case of CD14, CD21 and CD26; the starting conformations in t1, t2 and t3 are not detected during the MD calculations; however, there are highly rated conformations in every individual trajectory. The result suggests that, although MD calculations do not necessarily get stuck in the starting conformations, every trajectory cover areas that preferably surround certain positions.

Trajectory	**1**	
CD28	**name D-II**	**w (%)**
Starting Conf.	**3Z,S,+,Z,S,Z,a,2Z,S,2Z,2S,a,-,Z,S,-,2Z,S,a,+,a,-**	(-)
MD Conf.	**2Z,S,Z,-,2S,a,-,S,2Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a**	11.80
	2Z,S,Z,-,+,S,a,-,S,2Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	9.90
	2Z,S,Z,-,2S,a,-,S,-,Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	7.10
	2Z,S,Z,-,+,S,a,-,S,-,Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	6.60
	3Z,2S,Z,S,-,a,2Z,S,Z,-,+,S,a,-,S,2Z,S,-,S,a,+,a,-	5.30
	3Z,2S,Z,S,-,a,2Z,S,Z,-,2S,a,-,S,2Z,S,-,S,a,+,a,-	4.50
	3Z,2S,Z,S,-,a,2Z,S,Z,-,+,S,a,-,S,-,Z,S,-,S,a,+,a,-	3.80
	3Z,2S,Z,S,-,a,2Z,S,Z,-,2S,a,-,S,-,Z,S,-,S,a,+,a,-	2.90
TOTAL (%)	**8/253**	**51.90**

Trajectory	2	
CD28	name D-II	w (%)
Starting Conf.	5Z,a,-,Z,S,Z,S,+,Z,S,-,a,2Z,S,Z,2S,-,a,2S,-,a	(-)
MD Conf.	5Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a	11.40
	4Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a,-	8.10
	4Z,a,2Z,S,Z,4S,-,a,-,S,Z,2S,Z,-,a,2S,-,a,-	6.20
	4Z,a,2Z,S,Z,4S,-,+,-,S,Z,2S,Z,-,a,2S,-,a,-	5.00
	5Z,a,S,-,S,Z,4S,-,3S,Z,2S,2-,a,+,S,Z,a	3.70
	2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a,Z,-,S,2Z,a,S	3.00
	5Z,a,S,-,S,Z,4S,-,S,Z,S,Z,2S,2-,a,+,S,Z,a	2.70
	4Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,3S,a,-	2.60
TOTAL (%)	8/212	42.70

Trajectory	3	
CD28	name D-II	w (%)
Starting Conf.	2Z,S,2Z,S,-,a,Z,-,2S,Z,a,2S,-,S,Z,-,S,-,2S,Z,3S	(-)
MD Conf.	3Z,S,-,S,-,a,-,Z,2S,-,Z,2S,3Z,-,Z,2S,Z,S,Z,2S	16.70
	3Z,-,Z,2S,Z,S,Z,3S,2Z,S,-,S,-,a,-,Z,2S,-,Z,2S	8.40
	3Z,S,-,S,-,a,-,Z,2S,2-,2S,3Z,-,Z,2S,Z,S,Z,2S	6.80
	3Z,S,-,S,-,a,-,Z,S,Z,-,Z,2S,3Z,-,Z,2S,Z,S,Z,2S	5.30
	3Z,-,Z,2S,Z,S,Z,3S,2Z,S,-,S,-,a,-,Z,2S,2-,2S	3.70
	3Z,S,-,S,-,a,-,Z,2S,-,Z,2S,3Z,-,Z,2S,-,S,Z,2S	3.70
	4Z,-,Z,2S,Z,S,Z,2S,3Z,S,-,S,-,a,-,Z,2S,-,Z,S	2.70
	3Z,-,Z,2S,-,S,Z,3S,2Z,S,-,S,-,a,-,Z,2S,-,Z,2S	2.00
TOTAL (%)	8/256	49.30

Trajectory	average	
CD28	name D-II	w (%)
Starting Conf.		
MD Conf.	3Z,S,-,S,-,a,-,Z,2S,-,Z,2S,3Z,-,Z,2S,Z,S,Z,2S	5.57
	2Z,S,Z,-,2S,a,-,S,2Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	3.93
	5Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a	3.80
	2Z,S,Z,-,+,S,a,-,S,2Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	3.30
	3Z,-,Z,2S,Z,S,Z,3S,2Z,S,-,S,-,a,-,Z,2S,-,Z,2S	2.80
	4Z,a,S,2Z,5S,-,S,Z,S,Z,2S,2-,a,2S,Z,a,-	2.70
	2Z,S,Z,-,2S,a,-,S,-,Z,S,-,S,a,+,a,-,2Z,-,2S,Z,S,-,a	2.37
	3Z,S,-,S,-,a,-,Z,2S,2-,2S,3Z,-,Z,2S,Z,S,Z,2S	2.27
TOTAL (%)	8/721	26.73

Figure 6.37: The most important conformations from CD28 “in vacuo” MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

6.2.9.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 6.38] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 6.39]. The *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD28	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	253				
	2	212	721	721	240.33	**0.00**
	3	256				

Figure 6.38: *Lambda index* –trajectories overlapping ratio- for CD28.

- The three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 6.39]. Besides, the saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 6.38]; therefore, calculations are not trustworthy: on the one hand, it is not sure that trajectories fully converge; on the second hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturates, information about conformational space is unreliable.
- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is arguably underexplored.

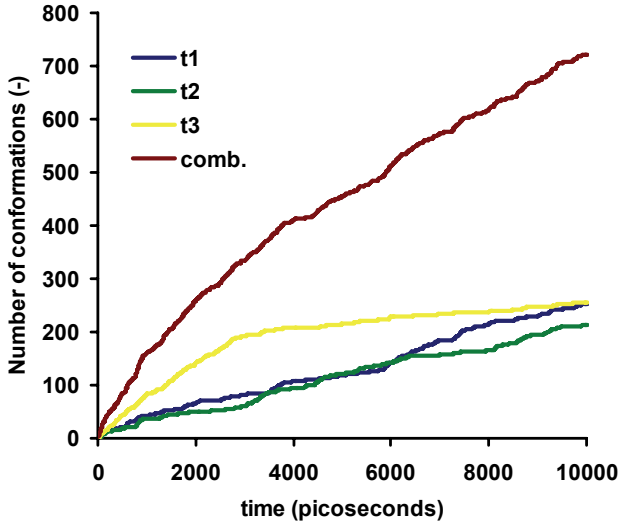


Figure 6.39: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD28.

The three trajectories display a similar efficiency in the individual exploration of conformational space: the general tendency is t1 ≈ t2 ≈ t3, although at the very beginning, t3 seems to grow faster than t1 or t2. Regarding the combined trajectory, it can be seen that it has nearly three times the number of conformations found in t1, t2 or t3; this means that every individual trajectory covered disconnected areas in conformational space, which, in addition, explains the 0% lambda index ratio.

Similarly to CD14, CD21 and CD26, the table in [Fig. 6.37] shows that –as the number of glucoses grows- the two-group behaviour vanishes and the *conformational noise* grows. The Pareto diagram [Fig. 6.40], again as happened in the cases of CD14, CD21 and CD26, seems to suggest that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, and similar to the CD21 and CD26 cases:

CD28		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	253	212	256	721
20% of Conf.	(n)	51	42	51	144
Acc. Stat. Weight	(%)	**76.00**	**75.40**	**74.20**	**75.30**

Figure 6.40: Pareto. CD28 MD trajectories in gas phase.

The best 20% conformations represent between 74% and 76% of their trajectories which is in agreement with the results explained in [Fig. 6.37] where the first 8 conformations of t1, t2, t3, and the combined trajectories were shown with their statistical weights: 52%, 43%, 49% and 27%. These ratios –surprisingly higher than those found in CD21 and CD26- suggest that they could still be considered as the *main conformations* –in the individual trajectories- while the others could be the *conformational noise*. Nevertheless, as was said in the previous cases of the large cyclodextrins, the noise is ruling the conformational behaviour.

6.2.9.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- suggests that the three trajectories are completely unconnected.

6.2.9.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

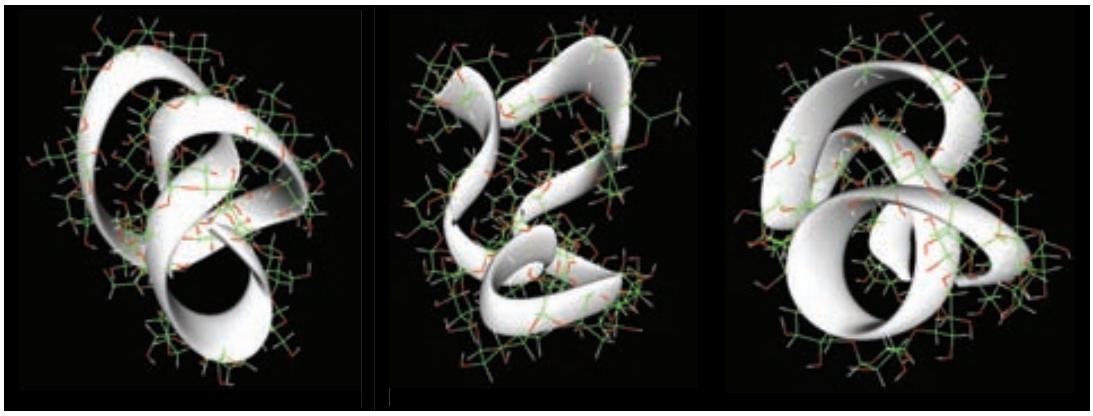


Figure 6.41: Average conformations from CD28 "in vacuo" MD trajectories t1, t2 and t3.

The situation is similar to that in CD21 and CD26. The images of the average structures [Fig. 6.41] contradict the results obtained from the descriptor analysis concluding that D-II is extremely sensitive to small conformational changes and therefore inadequate to describe CD28. Furthermore, Molecular Dynamics trajectories tend to remain close to the starting conformations.

- Under the new criterion, the three of them; t1, t2 and t3 can be regarded as different conformations. CD28 folds creating between two and three loops although not clearly seen because of the entangled conformations –in t1 and t3-. On the other hand, a loose and partially extended conformation with only one loop is obtained –in t2-.
- Hydrogen bonds are very important to stabilise conformations. Different types of arrangements coexist at the same time to allow loops, chains and helices because of the secondary structure.
- The directionality involving hydrogen bonds is a less important phenomenon not fully used in the family of the large CDs. However, partial sequences are detected.

6.3 GROUP PROPERTIES I: The Effect of the Set of Charges in CD5

Up to this point the properties and results for all the cyclodextrins examined have been –more or less- individually explained. However, sporadic comparisons involving

homologous calculations were included. Now, an important comparative study is specifically discussed within the current section involving groups of cyclodextrins: the effect of the charges in the Molecular Dynamics in CD5.

6.3.1 The AMBER philosophy: The “unit”

There are several software packages available in the market for Molecular Mechanics calculations. One of the most widely employed is maybe AMBER. This program is mostly oriented to protein calculations and therefore it comes with libraries including a variety of aminoacid residues.

Problems start when the user is interested in molecules other than proteins. Since AMBER compulsory urges the user to employ the residue philosophy prior to building the molecule, you are forced to create your own library, containing the necessary residues.

Creating a *unit*[199] –that is how a residue is called in AMBER terminology- is a tedious and unfriendly job involving several sequential stages –series of *ab initio* calculations and subsequent treatments- that ends furnishing the *prep* file containing the unit formatted to be read by LEaP module.

Briefly, this file contains topological information from two sources: the Force Field (See appendix in chapter IX), and the atom charges. The Force Field constants –different sets are available- are supplied by the software developers and are included during the unit creation steps just by “selecting” your Force Field preferences. The charges, however, are entirely left at the researcher criterion –although the best choice is always calculating the charges using the same conditions employed for the Force Field parameterisation- and therefore everybody is allowed to develop his/her own set of charges for his/her units.

[199] The process is explained in detail not only in the AMBER 7 manual but also in the theses of former group members: Dr. Iván Beà, Dr. Itziar Maestre, Dr. Miguel de Federico, and Dr. Javier Pérez. The last one also wrote a friendly Tutorial for beginners in AMBER.

This process was widely used in our group until the former members Dr. Miguel de Federico and Dr. Javier Pérez detected certain deficiencies in RESP that reflected on artefacts and non-sense sets of charges.

This result made us thinking about the possible effects and implications of this situation in further calculations when people using AMBER all around the world employ different –maybe *wrong?*- sets of charges for similar residues just because every one is allowed to do it under no supervision.

To evaluate the effect of the charges we designed an experiment: Firstly, the idea was creating two identical molecules that only differ in their sets of charges. Secondly, study them exactly under the same computational conditions and then compare the results.

6.3.2 Comparative Molecular Dynamics

In this sense, our choice was the CD5 cyclodextrin –the one with 5 glucoses- and two different sets of charges were used to create the glucose unit (see chapter XI appendix in the DVD), and then to build the macromolecule:

1) CD5i: Is the cyclodextrin generically called CD5 in this work. It carries the set of charges developed by Dr. Ivan Beà, and later incorporated by Dr. Itziar Maestre and Dr. Miguel de Federico.
2) CD5x: Is the set of charges developed by Dr. Javier Pérez during his Master and was created strictly following the AMBER –and specifically- RESP manuals; averaging statistically meaningful conformations. It was already mentioned that this is basically the *"different"* set of charges because all the CDs in this study –the whole family ranging from 5 to 28 glucoses- have been created with the other set of charges.

The two series of two Molecular Dynamics in gas phase [Fig. 6.42] were carried out exactly under the same conditions [*chapter VI, sections 6.2.1 and 6.2.9*]: same starting conformations, same number of trajectories, same thermal conditions, same length, same frequency sampling, same coupling… (The C-shell scripts and AMBER input

files were also the same). The only difference was just the set of charges in every cyclodextrin.

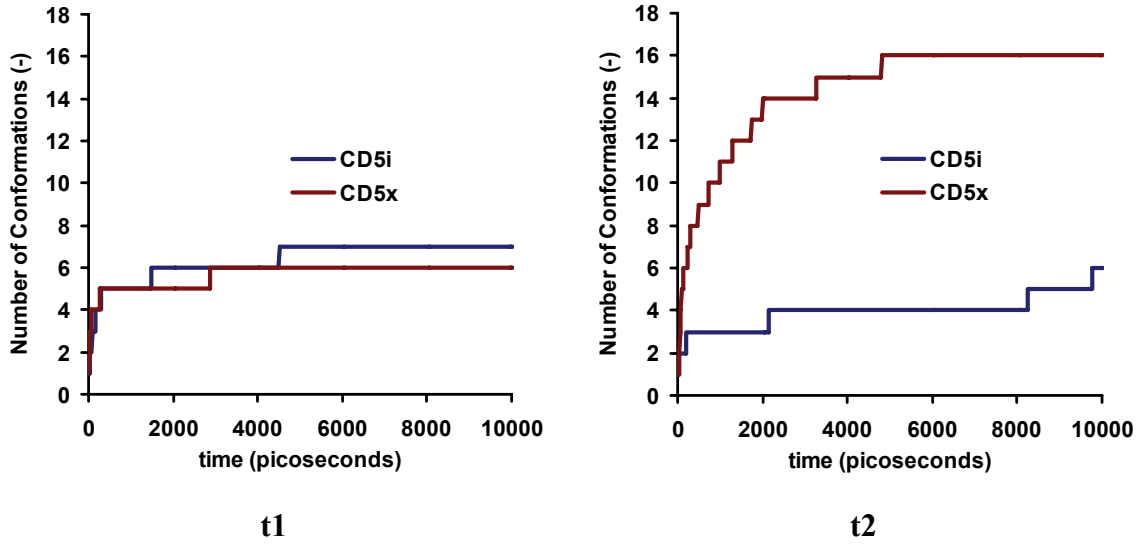


Figure 6.42: Number of conformations for CD5 in *t1* and *t2* obtained using the two different sets of charges; those employed by Dr. Ivan Beà, Dr. Itziar Maestre and Dr. Miguel de Federico (CD5i) and those developed by Dr. Javier Pérez (CD5x).

- D-II was the appropriate descriptor to characterize the family of the small CDs so the comparisons are based on the grounds of this descriptor.
- The tendency [Fig. 6.42] points to CD5x as the one that better explores the conformational space, although the situation is not so clear in t1, where both CD5i and CD5x behaves similarly.
- The same tendency is observed examining the evolution of the total number of conformations in the combined trajectory versus time [Fig. 6.43]. Since CD5x and CD5i conformational spaces have been successfully explored considering both t1 and t2 together, this result can be regarded as trustworthy.
- The same effect in flexibility –being CD5x more flexible than CD5i- was already detected when the Simulated Annealing calculations were analysed in Chapter V. It was then concluded that the difference in charges were of importance in the conformational behaviour.

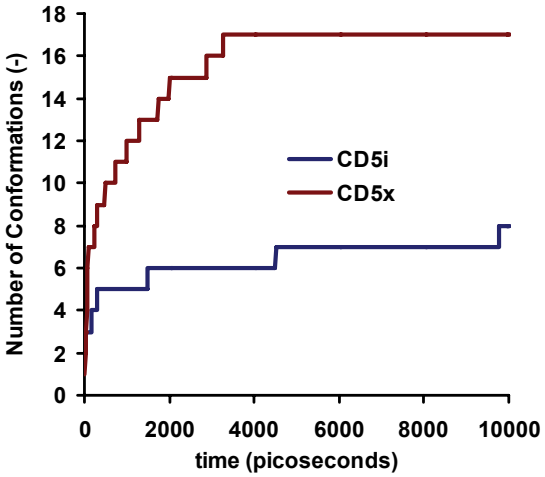


Figure 6.43: Number of total conformations for CD5 in the combined trajectories obtained using the two different sets of charges; those employed by Dr. Ivan Beà, Dr. Itziar Maestre and Dr. Miguel de Federico (CD5i) and those developed by Dr. Javier Pérez (CD5x).

6.3.3 Results

- The difference in charges influences on the conformational space and affects the calculations –at least, Simulated Annealing and Molecular Dynamics-.
- Two caveats: on the one hand, it has been established that systems are quite sensitive to alterations in charges. On the other hand, researchers are allowed to develop their own sets of charges with absolute freedom and under no supervision. This combination seems to be discouraging.

6.4 SUMMARY OF RESULTS

This section contains the most important results described in this chapter:

Regarding Methodology:

- Trajectory Overlapping Analysis –involving *omega* and specially *lambda* indexes- in combination with the Total Number of Conformations Analysis has proved to be a powerful MD conformational search tool for determining the Saturation when exploring the Conformational Space.
- It has been detected a MD dependence on starting conformations, meaning that MD trajectories tend to explore areas close to the starting position. This effect

remains almost undetected in small CDs but not in large ones, suggesting that it is timescale problem.

Regarding Cyclodextrins:

- According to flexibility, CDs can be divided in two groups on the basis of D-II:
  - Rigid and Semirigid [Fig. 6.44a]: Well defined conformations:
    - CD5x, CD5 and CD6: *Main conformations*.
    - CD7 and CD8: *Main conformations* and *Conformational Noise*.
  - Flexible [Fig. 6.44b]: No defined conformations:
    - CD10 and larger: *Conformational Noise*.

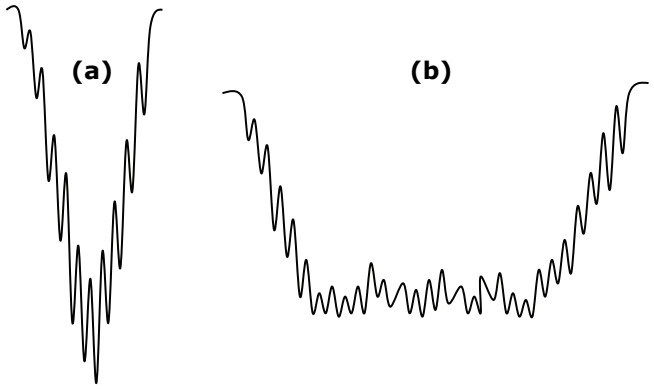
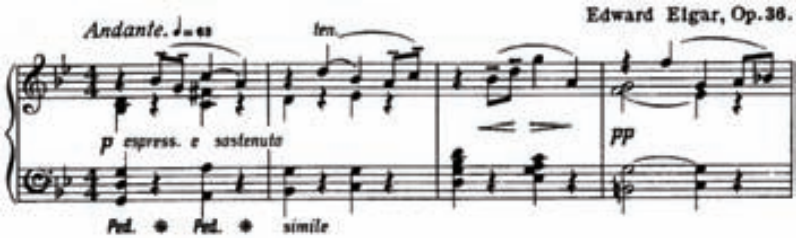


Figure 6.44: Schema of the conformational spaces for (a) small and (b) large cyclodextrins.

- Flexibility grows as the number of glucoses grows.
- Descriptor D-II is a good choice for small cyclodextrins. However, it is not the appropriate one for large CDs because it is unable to model folding phenomena.
- It is necessary to develop a new descriptor –more general and less precise- that correctly describes folding and the secondary structure.
- No conformation with inverted glucoses was found in CD5, CD5x and CD6 due to its extremely high rigidity.
- The use of different sets of charges in the molecules strongly affects the conformational behaviour of the system.



Sir Edward Elgar
Enigma Variations op. 36
(1899)
Theme "Enigma"

"All truths are easy to understand once they are discovered; the point is to discover them."

Galileo Galilei

7 RESULTS IV: MOLECULAR DYNAMICS IN SOLVENT

7.1 MOLECULAR DYNAMICS

This chapter is the continuation and final part of the research developed in chapter VI. Both of them can be considered as a single entity containing the Molecular Dynamics (MD) calculations. However, for an easier approach, it was decided to split the information into two parts: the first one containing the MD calculations in gas phase, and the second one containing the calculations in real solvent:

- Gas Phase MD calculations (*previous chapter*): CD5, CD6, CD7, CD8, CD14, CD21, CD26, CD28 and CD5x.
- Water MD calculations (*current chapter*): CD8, CD14, CD21, CD26, and CD28.
- Benzene MD calculations (*current chapter*): CD26.

Just reminding the two objectives established in Chapter VI, they are included again:

- Testing the efficiency of the MD methodology in combination with Trajectory Overlapping Analysis to prove/false MD dependence on starting conformation.
- Since MD models the *real* behaviour of molecules in vacuo/solvent; exploration of the Conformational Space of the CDs selected as benchmark and doing a comparative study of their conformational properties.

For every cyclodextrin, several MD trajectories –starting from different conformations- have been obtained in order to measure the overlapping area in conformational space. Some of the small cyclodextrins –CD5x, CD5, CD6, and CD7- were discarded for the MD water calculations. This decision was made because the SA conformational search and the MD calculations in gas phase proved to be efficient enough to study their whole conformational space. In this sense, it was also assumed that solvent effects could be of special interest in calculations where folding is important, i.e., those involving large CDs. Therefore, this procedure helped us saving a considerable amount of computational time.

CD26 was the only cyclodextrin studied using the tentative benzene solvent model half-developed in this group. The point of this is comparing the results for benzene with

those previously obtained in gas phase and water to evaluate the influence of solvent polarity in the conformational space of cyclodextrins. Depending on the results, further studies including other cyclodextrins could be considered as advisable and worthy. It is worth mentioning that experimental solubility values for CDs from 6 to 39 glucose units in water are available[202] and support the fact that all of them are soluble. Furthermore, there are several studies for CD26 –including crystallographic ones[203]- involving also water solvent, which means that water is probably the solvent most widely employed when working with this macrorings.

Descriptors I and II have been calculated for the whole sets of MD trajectories although only the second one has been extensively used in the present Thesis. Complete information regarding descriptors I and II, Markov matrices, statistics and chain of conformations not included in this chapter can be found in the DVD attached to the back cover.

7.2 Molecular Dynamics in Water Solvent (TIP3P H_20)

As said, the small cyclodextrins –CD5x, CD5, CD6, and CD7- have been discarded for the MD calculations in water solvent and only CD8 and the large CDs have been studied.

The system is conceived as a truncated octahedral box where the cyclodextrin is placed in the middle, and the solvent –the explicit TIP3P water solvent model in our case- surrounds the molecule until it fills up the remaining space within the unit cell. The series were carried out following the schema in [Fig. 7.1] and the full calculation includes four steps.

[202] (a) Taira, H.; Nagase, H.; Endo, T.; Ueda, H.; *J. Incl. Phenom. Macrocycl. Chem*. **2006**. *56(1-2)*. 23-28. (b) Lambertsen Larsen, K.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *43(1-2)*. 1-13. (c) Ueda, H.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*. 53-56. (d) Ueda, H.; Wakisaka, M.; Nagase, H.; Takaha, T.; Okada, S.; *J. Incl. Phenom. Macrocycl. Chem*. **2002**. *44(1-4)*. 403-405.
[203] (a) Gessler, K.; Usón, I.; Takaha, T.; Krauss, N.; Smith, S.M.; Okada, S.; Sheldrick, G.M.; Saenger, W.; *Proc. Natl. Acad. Sci. USA*. **1999**. *96(8)*. 4246-4251. (b) Nimz, O.; Gessler, K.; Usón, I.; Laettig, S.; Welfle, H.; Sheldrick, G.M.; Saenger, W.; *Carbohydr. Res*. **2003**. *338(9)*. 977-986. (c) Nimz, O.; Gessler, K.; Usón, I.; Saenger, W.; *Carbohydr. Res*. **2001**. *336(2)*. 141-153.

- Step zero is a restricted molecular energy minimization. It is designed to remove –if necessary- any remaining residual tension generated during the solvation process. By default algorithms –the first ten steps of *steepest descent* followed by *conjugated gradient*- are employed, including up to 50,000 steps or by-default termination energy criterion.
- The first step is the *heating slope* and *equilibration* process under NVT conditions –*canonical ensemble*-: 300 picoseconds length, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, 298 K in the equilibrium and external constraints to avoid chair-to-boat conformational interchanges in glucoses. Coupling constants are modulated along the simulation to avoid *blow-up* terminations.
- The second step is the *equilibration* process under NPT conditions –*isothermal-isobaric ensemble*-: 300 picoseconds length, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, bath thermal coupling of 0.5 picoseconds, temperature 298 K in the equilibrium and pressure 1 atmosphere. External constraints are removed and coupling constants are kept unchanged along the simulation.
- The third step is the *sampling* process under NPT conditions –*isothermal-isobaric ensemble*-. Its length is 5,000 picoseconds, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, 298 K, pressure 1 atmosphere, bath thermal coupling of 0.5 picoseconds and sampling frequency of 1 snapshot per picosecond –the trajectory is stored in this step for further analysis-. External constraints –to avoid chair-to-boat conformational interchanges in glucoses- are removed and coupling constants are kept unchanged along the simulation.

Detailed information regarding explicit TIP3P water solvent model MD conditions and further computational aspects can be found in the DVD: annex 10.1.1.5 and its sub-annexes (chapter X AMBER 7 files).

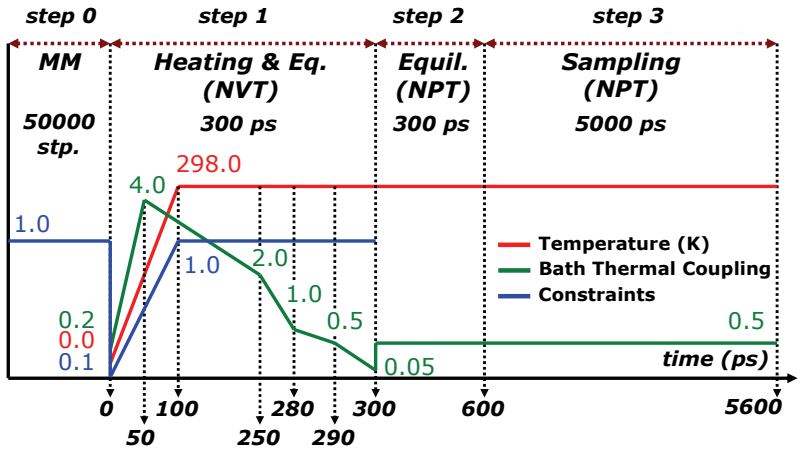


Figure 7.1: Schematic depiction of the 4-step water-solvent Molecular Dynamics protocol. Coloured lines, although not scaled, represent the dynamically coupled control parameters.

7.2.1 CD8 (γ-CD)

The set of Molecular Dynamics in water solvent carried out for CD8 includes four different starting conformations as was done with the homologous calculation in gas phase. Here again, the high flexibility in CD8 recommended this measure, which is also advisable if further comparisons with the CD8 calculations "in vacuo" are being done. The group of the starting conformations included: the first one [**Z,S,Z,S,Z,S,Z,S**]; the second one [**2Z,a,-,Z,S,Z,S**]; the third one [**2Z,2S,2Z,a,-**] and the fourth one [**3Z,2S,-,2S**]. All of them were selected according to statistical weight criteria and are the same as those employed in the gas phase MD calculations.

7.2.1.1 Population Analysis

The most representative results can be found in [Fig. 7.2]. In this case, 14 conformations have been selected and explain, in all cases, over 89% of the system. Some important conclusions derived from the table are shown hereafter:

- A total number of 444 conformations were found.

- The trajectories t1, t2, t3, t4 and the combined trajectory share in common only a few of the most important conformations, not necessarily located in the first places, in all the trajectories, all at the same time.
- The total number of conformations in t1, t2, t3, t4 and the combined trajectory is considerably different. In this sense, t1 and t4 present a similar number of conformations –in fact 125 and 150, respectively- and so does the group of t2 and t3 –with 267 and 254 conformations each-, while the total number of conformations in the combined trajectory –a total of 444 conformations- is considerably bigger in comparison with the individual trajectories.
- The most highly rated conformations in t1, t2, t3, t4 and the combined trajectory have experienced a dramatic reduction in their ratios, being not superior to 11%. This fact is specially noticeable in the combined trajectory where the weights of the three most important conformations; [**2Z,S,Z,S,-,2S**], [**Z,2S,-,2S,-,S**] and [**2S,-,2S,-,S,-**] are found in the range of 5%.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase, this result meaning that the solvent influence in the molecular behaviour is determinant.
- None of the starting conformations in t1, t2, t3 and t4 are visited again during the calculation time. The variety of possibilities seems to suggest that MD calculations do not necessarily get stuck in the starting conformations, although the partial overlapping suggests that every trajectory could cover areas that preferably surround the starting position.

Trajectory	1		2		3	
CD8	name D-II	w (%)	name D-II	w (%)	name D-II	w (%)
Starting Conf.	Z,S,Z,S,Z,S,Z,S	(-)	2Z,a,-,Z,S,Z,S	(-)	2Z,2S,2Z,a,-	(-)
MD Conf.	2S,-,2S,-,S,-	10.74	Z,S,Z,S,Z,S,-,S	5.66	2Z,S,Z,S,-,+,-	11.64
	Z,2S,-,2S,-,S	9.78	Z,S,-,S,Z,S,-,S	5.04	Z,S,-,2S,-,+,-	5.20
	2Z,S,Z,S,-,2S	9.34	2Z,2S,2Z,+,-	4.60	2S,-,S,-,+,2-	4.58
	Z,S,-,2S,-,2S	7.70	2Z,S,Z,S,-,2S	4.52	2Z,S,-,S,-,+,-	4.40
	2Z,2S,-,S,Z,S	7.66	2Z,2S,-,Z,+,-	4.12	Z,S,-,+,-,S,-,S	3.74
	Z,S,-,S,Z,S,-,S	5.56	2Z,2S,2Z,a,-	3.60	Z,+,2-,2S,-,S	3.50
	Z,S,Z,-,2S,-,S	4.90	Z,+,-,Z,-,2S,-	3.26	2Z,+,2-,S,Z,S	3.28
	Z,S,Z,S,Z,S,-,S	4.62	2Z,+,-,Z,-,2S	3.18	Z,S,-,+,-,Z,-,S	2.84
	Z,S,Z,S,-,2S,-	4.50	Z,2S,-,2S,-,S	3.16	Z,a,2-,2S,-,S	2.82
	Z,2S,-,2S,Z,S	3.14	2Z,2S,-,S,Z,S	2.88	2S,-,S,-,a,2-	2.20
	3Z,2S,-,2S	1.86	Z,S,-,2S,Z,-,S	2.82	2Z,S,Z,S,-,a,-	2.06
	2Z,S,-,S,-,2S	1.82	Z,S,-,2S,-,+,-	2.56	2Z,a,2-,S,Z,S	1.72
	2Z,S,2Z,S,Z,S	1.66	2S,-,2S,-,S,-	2.22	2Z,+,-,Z,S,Z,S	1.58
	Z,S,-,S,-,2S,-	1.54	2Z,a,-,Z,-,2S	2.10	Z,2S,-,+,-,Z,S	1.56
TOTAL (%)	14/125	74.82	14/267	49.72	14/254	51.12

Trajectory	4		average	
CD8	name D-II	w (%)	name D-II	w (%)
Starting Conf.	3Z,2S,-,2S			
MD Conf.	2Z,S,Z,S,-,2S	8.86	2Z,S,Z,S,-,2S	5.68
	Z,2S,-,2S,-,S	8.70	Z,2S,-,2S,-,S	5.41
	Z,S,-,2S,-,2S	7.70	2S,-,2S,-,S,-	5.08
	2S,-,2S,-,S,-	7.34	Z,S,-,2S,-,2S	4.27
	Z,S,Z,S,Z,S,-,S	5.86	Z,S,Z,S,Z,S,-,S	4.04
	2Z,2S,-,S,Z,S	5.12	2Z,2S,-,S,Z,S	3.92
	Z,S,-,S,Z,S,-,S	4.24	Z,S,-,S,Z,S,-,S	3.71
	2Z,S,2Z,S,Z,S	2.68	2Z,S,Z,S,-,+,-	3.14
	Z,S,Z,S,-,2S,-	2.62	Z,S,Z,S,-,2S,-	1.98
	2Z,2S,2Z,2S	2.60	Z,S,-,2S,-,+,-	1.94
	2Z,2S,Z,-,2S	2.54	Z,S,Z,-,2S,-,S	1.93
	Z,S,Z,S,-,S,-,S	2.34	Z,2S,-,2S,Z,S	1.65
	2Z,S,-,2S,-,S	2.34	Z,S,-,2S,Z,-,S	1.45
	Z,2S,-,2S,Z,S	2.22	2Z,2S,2Z,+,-	1.32
TOTAL (%)	14/150	65.16	14/444	45.50

Figure 7.2: Conformations from CD8 water MD trajectories 1, 2, 3, 4, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.2.1.2 Overlapping Analysis I: Accumulative Lambda Index

The *accumulative lambda index* [Fig. 7.3b] is calculated from the total number of conformations in t1, t2, t3, t4 and the combined trajectory [Fig. 7.3a].

- All four trajectories; t1, t2, t3 and t4, evolve in time towards better explorations of the conformational space, although none of them really saturates [Fig. 7.3a], at least, not under 4500 picoseconds. Nevertheless, as in the case of CD8 in gas phase, the individual and combined trajectories suggest that the system is close to a dynamical equilibrium. This point is confirmed by the saturation plot of lambda index vs. time, although the plot of total number of conformations vs. time is not fully converged.
- The total number of conformations in the combined trajectory and the lambda index saturate faster than the individual trajectories, so they can be regarded as good control parameters for monitoring the evolution of the system in time.
- The four individual trajectories t1, t2, t3 and t4 cannot separately explain the whole system; however, the combined information is still helpful in describing the conformational space.
- Trajectories t1, t2, t3 and t4 have in common about 59% of their conformational spaces –nearly 3 parts in 5- which is an almost identical rate to that obtained for CD8 in gas phase.

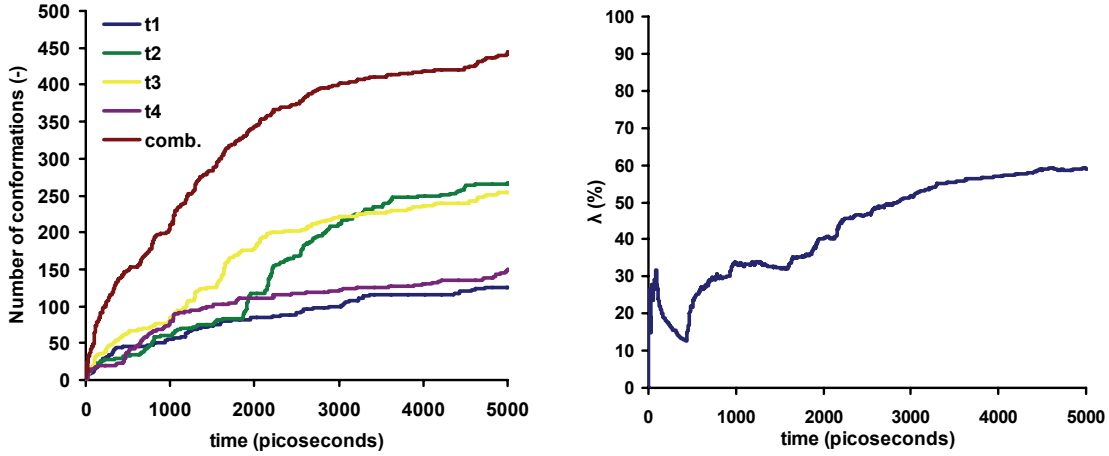


Figure 7.3: (a) Number of conformations in *t1, t2, t3, t4* and total number of conformations in the combined trajectory, *comb*. (b) *Accumulative Lambda index* –trajectories overlapping ratio- for CD8.

The efficiency in the individual exploration of the conformational space follows the order t1 ≈ t4 < t2 ≈ 3. The main remarkable phenomena in the series are: firstly, the area covered by t2 and t3 is wider than the area explored by t1 and t4; secondly, the smooth continuity observed in the exploration rate of t1 and t4, which grow almost in parallel; and finally, the noticeable increment in the number of conformations detected in t2 between 2000 and 4000 picoseconds. The last change suggests that the individual trajectory t2 probably discovered new areas in the conformational space not previously visited.

Furthermore, examining the table in [Fig. 7.2] under *Pareto's Law* [Fig. 7.4], it is found that the double behaviour –the main conformations and the conformational noise- is still there, although not so clearly observed:

CD8		Pareto				
		t1	t2	t3	t4	comb.
Number of Conf.	(n)	125	267	254	150	444
20% of Conf.	(n)	25	53	51	30	89
Acc. Stat. Weight	(%)	**87.44**	**83.04**	**79.98**	**84.96**	**86.13**

Figure 7.4: Pareto. CD8 MD trajectories in solvent water.

The best 20% conformations represent between 80% and 87% of their trajectories. The data in [Fig. 7.2] points to –more or less- the same information: the first 14

conformations of t1, t2, t3, t4 and the combined trajectories were shown with their statistical weights: 75%, 50%, 51%, 65% and 46%. These ratios suggest that they can be considered as the main conformations, while the others are the conformational noise.

7.2.1.3 Overlapping Analysis II: Omega Index

The omega index table [Fig. 7.5] shows that those conformations considered the main ones in t1, t2, t3 and t4 overlap and have high omega indexes. Poorly rated conformations –those that do not overlap in different trajectories- have omega values of 0% in some trajectories.

- The three most important conformations of CD8: [**2Z,S,Z,S,-,2S**] with 5.68%, [**Z,2S,-,2S,-,S**] with 5.41% and [**2S,-,2S,-,S,-**] with 5.08% experiment overlapping in rate between 30% and 48%.
- Many other not important conformations –statistically speaking- also overlap.
- The conformations with omega values of 0% -not shown in this table because of their low ratios- are those that have been detected in just one single trajectory but not in the other ones, all at the same time.

name D-II	t1	t2	t3	t4	$<{}^{i}p_j>$	${}^{i}p_{j(max)}$	${}^{i}p_{j(max)}/r$	${}^{i}\nu_j$	${}^{i}d_j$	${}^{i}\omega_j(\%)$
2Z,S,Z,S,-,2S	9.34	4.52	0.00	8.86	**5.68**	9.34	2.34	3.35	7.01	**47.75**
Z,2S,-,2S,-,S	9.78	3.16	0.00	8.70	**5.41**	9.78	2.45	2.97	7.34	**40.42**
2S,-,2S,-,S,-	10.74	2.22	0.00	7.34	**5.08**	10.74	2.69	2.39	8.06	**29.67**
Z,S,-,2S,-,2S	7.70	1.68	0.00	7.70	**4.27**	7.70	1.93	2.35	5.78	**40.61**
Z,S,Z,S,Z,S,-,S	4.62	5.66	0.00	5.86	**4.04**	5.86	1.47	2.57	4.40	**58.48**
2Z,2S,-,S,Z,S	7.66	2.88	0.00	5.12	**3.92**	7.66	1.92	2.00	5.75	**34.81**
Z,S,-,S,Z,S,-,S	5.56	5.04	0.00	4.24	**3.71**	5.56	1.39	2.32	4.17	**55.64**
2Z,S,Z,S,-,+,-	0.00	0.90	11.64	0.00	**3.14**	11.64	2.91	0.23	8.73	**2.58**
Z,S,Z,S,-,2S,-	4.50	0.80	0.00	2.62	**1.98**	4.50	1.13	0.86	3.38	**25.33**
Z,S,-,2S,-,+,-	0.00	2.56	5.20	0.00	**1.94**	5.20	1.30	0.64	3.90	**16.41**
Z,S,Z,-,2S,-,S	4.90	1.62	0.00	1.18	**1.93**	4.90	1.23	0.70	3.68	**19.05**
Z,2S,-,2S,Z,S	3.14	1.24	0.00	2.22	**1.65**	3.14	0.79	0.87	2.36	**36.73**
Z,S,-,2S,Z,-,S	1.30	2.82	0.00	1.68	**1.45**	2.82	0.71	0.75	2.12	**35.22**
2Z,2S,2Z,+,-	0.00	4.60	0.68	0.00	**1.32**	4.60	1.15	0.17	3.45	**4.93**
2Z,S,-,2S,-,S	1.46	1.44	0.00	2.34	**1.31**	2.34	0.59	0.73	1.76	**41.31**
Z,S,Z,S,-,S,-,S	1.50	1.34	0.04	2.34	**1.31**	2.34	0.59	0.72	1.76	**41.03**
2S,-,S,-,+,2-	0.00	0.64	4.58	0.00	**1.31**	4.58	1.15	0.16	3.44	**4.66**
2Z,S,2Z,S,Z,S	1.66	0.50	0.00	2.68	**1.21**	2.68	0.67	0.54	2.01	**26.87**
Z,S,Z,S,Z,S,Z,S	1.44	1.44	0.00	1.92	**1.20**	1.92	0.48	0.72	1.44	**50.00**
2Z,S,-,S,-,+,-	0.02	0.28	4.40	0.00	**1.18**	4.40	1.10	0.08	3.30	**2.27**
2Z,2S,-,Z,+,-	0.00	4.12	0.26	0.00	**1.10**	4.12	1.03	0.06	3.09	**2.10**
2Z,2S,2Z,a,-	0.00	3.60	0.60	0.00	**1.05**	3.60	0.90	0.15	2.70	**5.56**
TOTAL (%)					**55.15**					

Figure 7.5: Omega index –individual conformations overlapping ratio- for CD8.

7.2.1.4 Average Structures

CD8 belongs to the group of the small cyclodextrins so the ordinary view is employed:

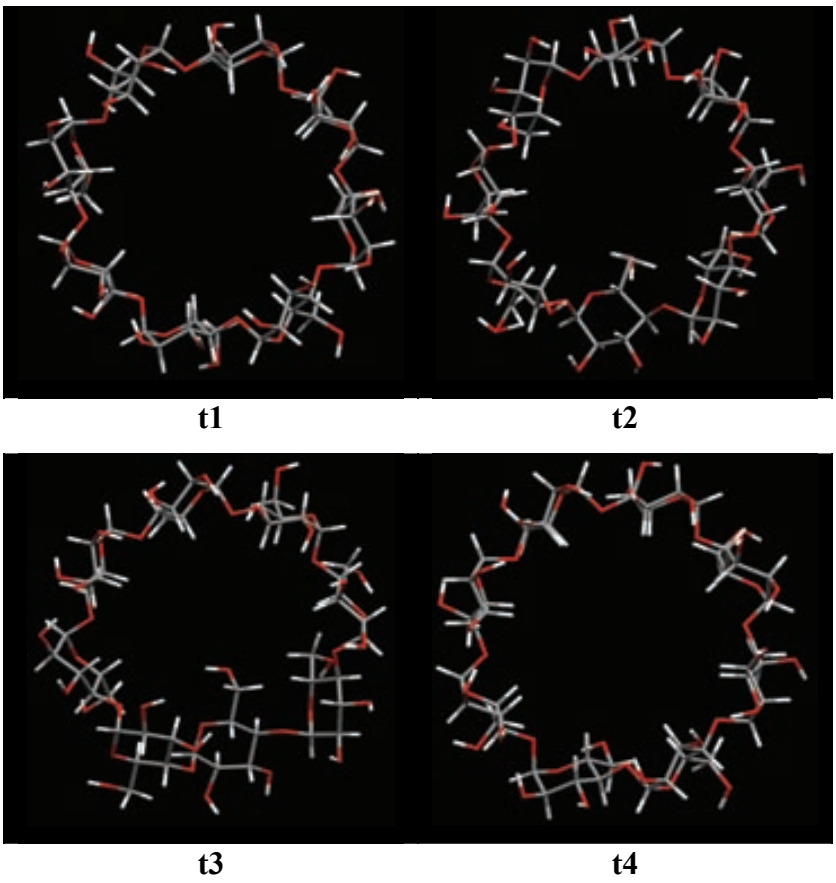


Figure 7.6: Average conformations from CD8 water MD trajectories t1, t2, t3 and t4.

The images [7.6] support the results obtained from the descriptor analysis:

- The average structures from t1, t2, t3 and t4 are viable molecules.
- It can be seen that the cavity is clearly maintained, a consequence of the remaining rigidity of CD8.
- The long chains of hydrogen bonds in CD8 no longer involve the full molecules. The effect of water solvent is important now, promoting other stable conformations where glucoses arrange their relative positions in [+] and [-], meaning that their hydroxyls mainly interact with solvent.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms –less important now than in the CD8 gas phase calculations- cannot be

detected by D-I or D-II; however, the overall behaviour of the macrocycle is still well described.

- The cyclodextrins are stabilized both by intramolecular –the long chains of hydrogen bonds- and also by intermolecular –solvent solute- interactions.

7.2.2 CD14

CD14 is one of the four cyclodextrins –alongside with CD21, CD26 and CD28- that belongs to the group of the *Large Cyclodextrins*. The extreme flexibility –already detected during the Simulated Annealing and the gas phase Molecular Dynamics calculations- is a determinant factor in their conformational behaviour.

The series of SA conformational search calculations carried out on this group of cyclodextrins proved that there were no preferred conformations in the ensembles but, mainly, *molecular noise*. In this sense, the series of MD calculations in gas phase not only agreed with this result but also showed a noticeable increase in the *molecular noise* at the expense of the *main conformations*. The new series of MD calculations in water solvent for large cyclodextrins were planned as a way to evaluate the effect of solvent in this group.

The set of Molecular Dynamics carried out for CD14 in water solvent included three different starting conformations: the first one [**2Z,S,Z,S,Z,2S,2Z,S,-,2S**]; the second one [**3Z,S,Z,S,Z,S,Z,S,Z,S,Z,S**] and the third one [**3Z,-,S,Z,2S,Z,a,Z,3S**]. All of them were randomly selected and are the same as those employed in the gas phase MD calculations.

7.2.2.1 Population Analysis

Some of the most representative results of the population analysis can be found in [Fig. 7.7]. As was done in the case of CD14 in gas phase, 8 conformations have been selected, explaining between 18% and 40% of the system, which is quite a difference in comparison with the rates measured in the gas phase calculation –between 38% and 71%-. This is not only a rather poor result, but also a worse result than that one

measured in the series of gas phase MD calculations. Some conclusions are shown hereafter:

- A total number of 1544 conformations were found.
- The total number of conformations found in the individual trajectories t1, t2 and t3 is reasonably different –respectively 471, 386 and 687 conformations- however, the total number of conformations in the combined trajectory is nearly three times bigger –about 1544-.
- The trajectories t1, t2, t3 and the combined trajectory do not have in common any of the conformations. In fact, none of them overlaps.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**2Z,S,-,S,-,S,Z,+,-,2S,-,S**], [**Z,3S,Z,S,Z,S,Z,S,-,S,Z,S**], [**Z,2S,-,Z,S,2-,2S,Z,S,-,a**] and [**Z,3S,Z,S,Z,S,Z,S,-,S,Z,S**]. Nevertheless, their ratios are lower than those measured for the small cyclodextrins.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase, this result meaning that the solvent influence on the molecular behaviour is determinant.
- The starting conformations in t1, t2 and t3 are not detected during the MD calculations; however, there are highly rated conformations in every individual trajectory. The result suggests that, on the one hand, MD calculations do not necessarily get stuck in the starting conformations, and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1		2	
CD14	name D-II	w (%)	name D-II	w (%)
Starting Conf.	2Z,S,Z,S,Z,2S,2Z,S,-,2S	(-)	3Z,S,Z,S,Z,S,Z,S,Z,S,Z,S	(-)
MD Conf.	2Z,S,-,S,-,S,Z,+,-,2S,-,S	8.68	Z,3S,Z,S,Z,S,Z,S,-,S,Z,S	11.22
	4Z,S,-,S,-,S,Z,+,-,2S	5.94	Z,3S,Z,S,Z,S,-,S,-,S,Z,S	6.46
	2Z,S,-,S,-,S,Z,+,-,2S,Z,S	5.28	Z,3S,Z,S,Z,S,-,S,Z,S,Z,S	6.10
	4Z,S,-,S,Z,S,Z,+,-,2S	3.38	Z,3S,-,S,Z,S,Z,S,-,S,Z,S	4.84
	3Z,3S,-,3S,3Z,S	3.04	Z,3S,-,S,Z,S,-,S,-,S,Z,S	3.68
	4Z,S,Z,-,3S,-,3S	2.62	Z,3S,Z,S,-,S,Z,S,-,S,Z,S	2.96
	4Z,2S,-,3S,2Z,-,S	2.48	Z,3S,Z,S,Z,S,-,S,-,S,-,S	2.70
	3Z,3S,-,S,Z,S,3Z,S	2.10	Z,3S,Z,S,Z,S,-,S,Z,S,-,S	2.20
TOTAL (%)	8/471	33.52	8/386	40.16

Trajectory	3		average	
CD14	name D-II	w (%)	name D-II	w (%)
Starting Conf.	3Z,-,S,Z,2S,Z,a,Z,3S	(-)		
MD Conf.	Z,2S,-,Z,S,2-,2S,Z,S,-,a	6.12	Z,3S,Z,S,Z,S,Z,S,-,S,Z,S	3.74
	2Z,S,2-,3S,-,S,a,-,2S	3.42	2Z,S,-,S,-,S,Z,+,-,2S,-,S	2.89
	Z,2S,-,Z,S,2-,2S,-,S,Z,a	3.26	Z,3S,Z,S,Z,S,-,S,-,S,Z,S	2.15
	Z,S,2-,3S,-,S,a,-,2S,-	2.92	Z,2S,-,Z,S,2-,2S,Z,S,-,a	2.04
	Z,2S,-,Z,S,2-,2S,Z,S,Z,a	2.78	Z,3S,Z,S,Z,S,-,S,Z,S,Z,S	2.03
	4Z,-,2S,Z,S,-,a,Z,2S	2.76	4Z,S,-,S,-,S,Z,+,-,2S	1.98
	Z,2S,-,Z,S,2-,S,Z,2S,-,a	2.60	2Z,S,-,S,-,S,Z,+,-,2S,Z,S	1.76
	2Z,S,2-,2S,-,S,Z,a,Z,2S	2.32	Z,3S,-,S,Z,S,Z,S,-,S,Z,S	1.61
TOTAL (%)	8/687	26.18	8/1544	18.21

Figure 7.7: Conformations from CD14 water MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.2.2.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 7.8] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 7.9]. The *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD14	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	471	1544	1544	514.67	**0.00**
	2	386				
	3	687				

Figure 7.8: *Lambda index* –trajectories overlapping ratio- for CD14 at the end of the MD calculations in water.

- The three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 7.9]. Moreover, the tendency no longer seems to be a saturation function but a potential/exponential one; meaning that the conformational space has just begun to be explored. Besides, the saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 7.8], therefore, calculations are not trustworthy: on the one hand, trajectories do not converge; on the other hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. The parameters, anyway, are still good ones for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since

none of them saturates and the tendency is nearly exponential, information about the conformational space, in the best, is unreliable.

- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Therefore, the system is underexplored.

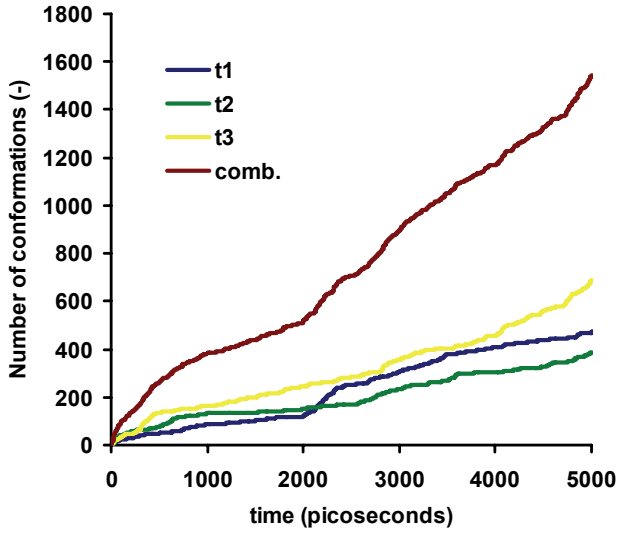


Figure 7.9: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD14 in water.

The efficiency in the individual exploration of the conformational space follows the order t3 > t1 > t2, although the individual differences in the number of conformations –about 100 ones- is reasonable in comparison with the total number of conformations in the combined trajectory –the ratio is about 1 out of 15-. The main remarkable phenomena in the series are; firstly, the tendency change in curvature –from saturation to exponential/potential-; and secondly, the smooth continuity observed in the exploration rates of t1, t2 and t3, that grow almost in parallel.

Furthermore, examining the table in [Fig. 7.7] under *Pareto's Law* [Fig. 7.10], it is found that the double behaviour –the main conformations and the conformational noise- is still there although not so clearly observed:

CD14		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	471	386	687	1544
20% of Conf.	(n)	94	77	137	309
Acc. Stat. Weight	(%)	81.98	83.14	79.64	81.97

Figure 7.10: Pareto. CD14 MD trajectories in solvent water.

The best 20% conformations represent between 80% and 83% of their trajectories. The data in [Fig. 7.7] points to –more or less- the same information: the first 8 conformations of t1, t2, t3 and the combined trajectories were shown with their statistical weights: 34%, 40%, 26% and 18%. These ratios are still good enough to support that they can be considered as the main conformations, while the others are the conformational noise.

7.2.2.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- clearly states that the three trajectories are completely unconnected.

7.2.2.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

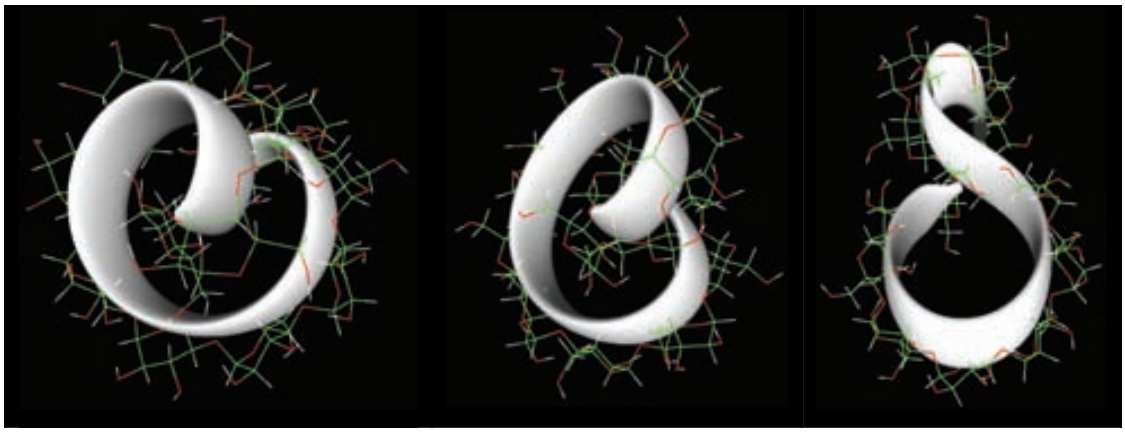


Figure 7.11: Average conformations from CD14 water MD trajectories t1, t2 and t3.

The images in [Fig. 7.11], as happened in the gas phase calculations, contradict the results obtained from the D-II descriptor analysis. The only difference is found when comparing the structures: t1, t2 and t3 were almost "identical" in the gas phase calculation, whereas, in the water solvent MD calculations, t1 and t2 are similar –and

similar to those obtained "in vacuo"- but t3 is different. The small loop in t3 is placed out of the cavity.

- The average structures from t1, t2 and t3 –similarly to the case of the small cyclodextrins- are viable molecules.
- Under the new criterion, t1 and t2 can be regarded as similar conformations; folding with a small internal loop. However, t3 is different; with an outer loop, describing a sort of an "8"-shape.
- Intra and intermolecular Hydrogen bonds are very important to stabilise conformations. The changes in the Secondary Structure are the consequence of the new environment.
- The directionality is a less important phenomenon since solvent-solute interactions break the long chains of hydrogen bonds.

7.2.3 CD21

CD21 is another cyclodextrin that belongs to the group of the *Large Cyclodextrins*. As in the case of CD14, three conformations were selected for the MD calculations in water. The set of Molecular Dynamics carried out for CD21 included three different starting conformations: the first one [**5Z,S,2Z,2S,Z,-,Z,S,Z,S,Z,S,Z,2S**]; the second one [**2Z,a,-,Z,S,-,S,Z,S,Z,a,Z,3S,-,Z,S,-,S**] and the third one [**2Z,-,S,Z,2S,-,S,Z,S,Z,2S,a,2S,-,a,-,S**]. All of them were randomly selected and are the same as those employed in the gas phase MD calculations.

7.2.3.1 Population Analysis

The most representative results of the population analysis can be found in [Fig. 7.12]. As was done in the case of CD21 in gas phase, 8 conformations have been selected, explaining between 5% and 16% of the system, which is a significant difference in comparison with the rates measured in the gas phase calculation –between 21% and 46%-. The reduction in the ratios is not only a poor result, but also a worse result than that one measured in the series of gas phase MD calculations. Some conclusions are shown hereafter:

- A total number of 4150 conformations were found.
- The total number of conformations found in the individual trajectories t2 and t3 is similar –respectively 1103 and 1214-. Trajectory t1 is, however, noticeably wider with 1833. The total number of conformations found in the combined trajectory is nearly three times bigger –about 4150-.
- The trajectories t1, t2, t3 and the combined trajectory do not have in common any of the conformations. In fact, none of them overlaps.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**2Z,-,S,Z,2S,2Z,-,S,Z,S,-,S,Z,2S,-,S,-**], [**Z,2S,-,Z,S,Z,a,-,2S,Z,S,-,S,-,2S,Z,a,-**], [**2Z,2-,2S,-,S,-,3S,-,2S,a,Z,S,-,a,-**] and [**Z,2S,-,Z,S,Z,a,-,2S,Z,S,-,S,-,2S,Z,a,-**]. Nevertheless, their ratios are lower than those measured for the small cyclodextrins and the gas phase calculations.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase, this result meaning that the solvent influence in the molecular behaviour is an important one.
- The starting conformations in t1, t2 and t3 are not detected during the MD calculations. Furthermore, the highly rated conformations are so poorly weighted that they cannot be considered as important. The result suggests that, on the one hand, MD calculations do not necessarily get stuck in the starting conformations, and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1		2	
CD21	name D-II	w (%)	name D-II	w (%)
Starting Conf.	5Z,S,2Z,2S,Z,-,Z,S,Z,S,Z,S,Z,2S	(-)	2Z,a,-,Z,S,-,S,Z,S,Z,a,Z,3S,-,Z,S,-,S	(-)
MD Conf.	2Z,-,S,Z,2S,2Z,-,S,Z,S,-,S,Z,2S,-,S,-	1.42	Z,2S,-,Z,S,Z,a,-,2S,Z,S,-,S,-,2S,Z,a,-	2.60
	2Z,-,S,Z,S,-,S,Z,2S,Z,S,-,Z,S,-,S,Z,2S	1.32	2Z,a,-,Z,2S,-,Z,S,Z,a,-,+,S,Z,S,Z,-,2S	2.28
	2Z,-,S,Z,2S,2Z,-,S,Z,S,-,S,Z,2S,Z,S,-	1.32	Z,2S,-,Z,S,Z,+,-,2S,Z,S,-,S,-,2S,Z,a,-	2.04
	2Z,-,S,Z,S,-,S,Z,2S,-,S,-,Z,S,-,S,Z,2S	1.26	2Z,-,2S,Z,a,-,Z,2S,-,S,Z,S,+,-,2S,Z,S	1.94
	6Z,S,Z,2S,2Z,-,2S,-,2S,Z,2S	0.92	Z,2S,-,Z,S,Z,a,-,+,S,Z,S,-,S,-,S,Z,S,a,-	1.74
	5Z,S,2Z,2S,Z,2-,S,2Z,S,2Z,2S	0.72	Z,2S,-,Z,S,Z,a,-,+,S,Z,S,-,S,-,2S,Z,a,-	1.68
	3Z,2S,Z,S,2-,S,3Z,2S,-,2S,Z,2-	0.72	Z,2S,-,S,Z,S,+,-,+,S,Z,S,-,S,-,2S,Z,a,-	1.68
	2Z,2S,2Z,-,S,-,S,Z,-,S,Z,+,-,Z,-,S,Z,S	0.72	2Z,-,2S,Z,a,-,Z,2S,-,S,-,S,+,-,2S,Z,S	1.58
TOTAL (%)	8/1833	8.40	8/1103	15.54

Trajectory	3		average	
CD21	name D-II	w (%)	name D-II	w (%)
Starting Conf.	2Z,-,S,Z,2S,-,S,Z,S,Z,2S,a,2S,-,a,-,S	(-)		
MD Conf.	2Z,2-,2S,-,S,-,3S,-,2S,a,Z,S,-,a,-	2.36	Z,2S,-,Z,S,Z,a,-,2S,Z,S,-,S,-,2S,Z,a,-	0.87
	2Z,2-,2S,-,S,-,S,Z,S,Z,2S,a,S,Z,-,a,-	1.24	2Z,2-,2S,-,S,-,3S,-,2S,a,Z,S,-,a,-	0.79
	3Z,a,-,S,3-,2S,Z,S,-,3S,-,2S,a	1.08	2Z,a,-,Z,2S,-,Z,S,Z,a,-,+,S,Z,S,Z,-,2S	0.76
	3Z,a,-,S,-,Z,-,2S,Z,S,-,3S,-,2S,a	1.02	Z,2S,-,Z,S,Z,+,-,2S,Z,S,-,S,-,2S,Z,a,-	0.68
	Z,2S,a,S,-,Z,a,-,S,-,Z,-,2S,Z,S,-,3S	0.98	2Z,-,2S,Z,a,-,Z,2S,-,S,Z,S,+,-,2S,Z,S	0.65
	2Z,2-,2S,Z,S,-,3S,Z,2S,a,2Z,-,a,-	0.98	Z,2S,-,Z,S,Z,a,-,+,S,Z,S,-,S,-,S,Z,S,a,-	0.58
	2Z,-,a,-,S,3-,2S,Z,S,-,3S,-,2S,a	0.94	Z,2S,-,Z,S,Z,a,-,+,S,Z,S,-,S,-,2S,Z,a,-	0.56
	2Z,2-,2S,-,S,-,3S,-,2S,a,S,-,Z,a,-	0.94	Z,2S,-,S,Z,S,+,-,+,S,Z,S,-,S,-,2S,Z,a,-	0.56
TOTAL (%)	8/1214	9.54	8/4150	5.44

Figure 7.12: Conformations from CD21 water MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.2.3.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 7.13] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 7.14]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD21	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	1833	4150	4150	1383.33	**0.00**
	2	1103				
	3	1214				

Figure 7.13: *Lambda index* –trajectories overlapping ratio- for CD21 at the end of the MD calculations in water.

- All three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 7.14]. The saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 7.13]. Therefore, calculations are not trustworthy because; on the one hand, trajectories do not converge; on the second hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturates, information about the conformational space is unreliable.

- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is underexplored.

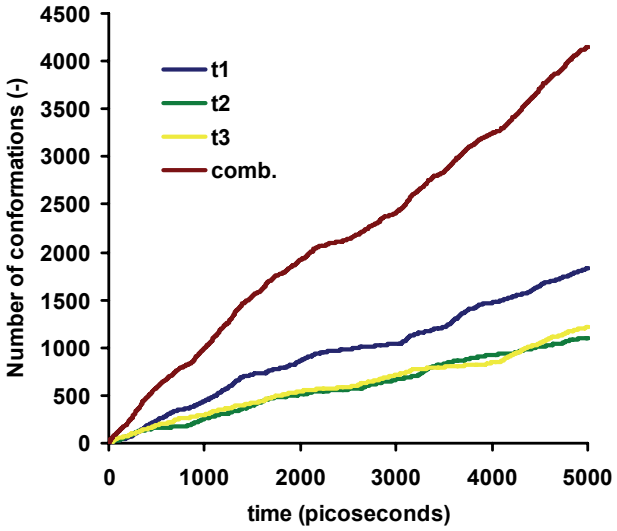


Figure 7.14: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD21 in water.

The efficiency in the individual exploration of the conformational space follows the order t1 > t2 ≈ t3. However, the total number of conformations in the combined trajectory is three times bigger, meaning that they cover separated areas. The main remarkable phenomena in the series are; on the one hand, the linear behaviour of the individual trajectories –that meet at zero steps- and on the other hand, the smooth continuity observed in the exploration rate of t1, t2 and t3 –where t2 and t3 grow almost in parallel-.

Examining the table in [Fig. 7.12] it is found that the two-group behaviour vanishes at the expense of the *conformational noise*. The Pareto diagram [Fig. 7.15] seems to suggest –from an optimistic point of view- that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, in fact, it is between 64% and 73%. This scenario is different from that one observed in CD14 in water solvent and also in CD21 in gas phase; in both cases the situation was slightly better than now.

CD21		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	1833	1103	1214	4150
20% of Conf.	(n)	367	221	243	830
Acc. Stat. Weight	(%)	**63.96**	**73.26**	**69.94**	**69.59**

Figure 7.15: Pareto. CD21 MD trajectories in solvent water.

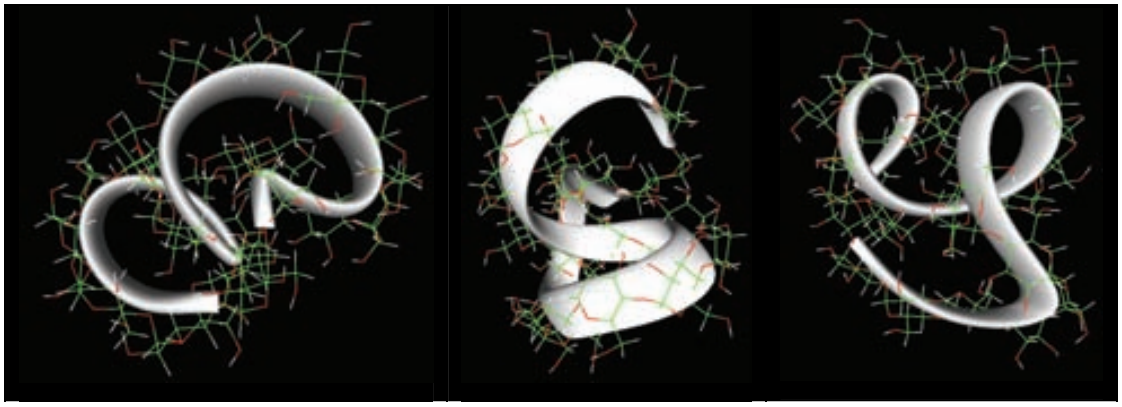
As said, the best 20% conformations represent between 64% and 73% of their trajectories. The data in [Fig. 7.12] points to –more or less- the same information: the first 8 conformations of t1, t2, t3 and the combined trajectories were shown with their statistical weights: 8%, 16%, 10% and 5%. These ratios are not good enough to support the view that they can be considered as the main conformations. The conformational noise is –step by step- starting to rule the general conformational behaviour.

7.2.3.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- clearly states that the three trajectories are completely unconnected.

7.2.3.4 Average Structures

CD21 is also one of the large cyclodextrins and then, protein-like visualizations of the type strands, ribbons and bands are employed:



t1 **t2** **t3**

Figure 7.16: Average conformations from CD21 water MD trajectories t1, t2 and t3.

Images in [Fig. 7.16] contradict the results obtained from the descriptor analysis. These results prove that, for our purposes, D-II is extremely sensitive to small conformational changes and therefore is inadequate to describe CD21. Likewise, Molecular Dynamics trajectories tend to be close to the starting conformations.

- Under the new criterion, the three of them; t1, t2 and t3 are also different conformations. In all three cases, CD21 folds creating double internal loops –in t1 and t2- and a tentative kind of helix –in t3-.
- Intra- and intermolecular hydrogen bonds are very important to stabilize the secondary structure.
- The directionality involving hydrogen bonds between hydroxyls and oxygen atoms is not fully observed in the family of the large CDs. Partial sequences are detected, however, the length of the chain rarely includes the total length of the macroring.

7.2.4 CD26

CD26 also belongs to the group of the *Large Cyclodextrins*. Similarly to what was done in the cases of CD14 and CD21, three conformations were selected for the MD calculations in water. The set of Molecular Dynamics carried out for CD26 included three different starting conformations: the first one [**3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S**]; the second one [**2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S**] and the third one [**3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a**]. All of them were randomly selected and are the same as those employed in the gas phase MD calculations.

7.2.4.1 Population Analysis

The most representative results of the population analysis can be found in [Fig. 7.17]. As was done in the case of CD21 in gas phase, 8 conformations have been selected, explaining between 6% and 18% of the system, which is a significant difference in comparison with the rates measured in the gas phase calculation –between 20% and 37%-. The reduction in the ratios is not a good result; in fact it is a worse result than that

one measured in the series of gas phase MD calculations. Some conclusions are shown hereafter:

- A total number of 5630 conformations were found.
- The total number of conformations found in the individual trajectories t1, t2 and t3 is reasonably different –respectively 1221, 1980 and 2429-. The total number of conformations found in the combined trajectory is nearly three times bigger –about 5630-.
- The trajectories t1, t2, t3 and the combined trajectory do not have in common any of the conformations. In fact, none of them overlaps.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,3Z,S**], [**2Z,S,2Z,2a,Z,2S,-,Z,2-,2S,-,2Z,a,2-,S,Z,S,a**], [**2Z,a,-,S,-,S,Z,-,3S,2-,S,-,Z,2S,-,+,3S,a,-**] and [**4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,3Z,S**]. Nevertheless, their ratios are lower than those measured for the small cyclodextrins and the gas phase calculations.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase, this result meaning that the solvent influence in the molecular behaviour is an important one.
- The starting conformations in t1, t2 and t3 are not detected during the MD calculations. Furthermore, the highly rated conformations are so poorly weighted that they cannot be considered as important. The result suggests that, on the one hand, MD calculations do not necessarily get stuck in the starting conformations, and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S	(-)
MD Conf.	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,3Z,S	3.34
	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,Z,2S,3Z,S	3.04
	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,2Z,-,S	2.44
	3Z,S,2Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,2Z,-,S	2.18
	3Z,S,2Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,Z,S,-,S	1.88
	3Z,S,-,Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,2Z,-,S	1.66
	3Z,S,2Z,2S,-,a,2Z,2S,2-,Z,+,-,3S,Z,S,-,S	1.66
	3Z,S,-,Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,Z,S,-,S	1.54
TOTAL (%)	8/1221	17.74

Trajectory	2	
CD26	name D-II	w (%)
Starting Conf.	2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S	(-)
MD Conf.	2Z,S,2Z,2a,Z,2S,-,Z,2-,2S,-,2Z,a,2-,S,Z,S,a	1.94
	2Z,S,2Z,2a,Z,2S,-,Z,-,Z,2S,-,2Z,a,2-,S,Z,S,a	1.26
	2Z,S,2Z,2a,Z,2S,-,Z,2-,2S,-,2Z,a,2-,S,Z,S,-	0.88
	2Z,S,-,S,a,+,-,2S,2Z,2-,S,Z,-,2Z,a,-,Z,S,-,S,a	0.80
	2Z,S,2Z,2a,-,2S,Z,2-,Z,2S,-,2Z,a,-,Z,S,Z,S,a	0.76
	2Z,S,2Z,2a,Z,2S,-,Z,-,Z,2S,-,2Z,a,-,Z,S,Z,S,a	0.72
	2Z,S,2Z,2a,Z,2S,-,Z,-,Z,2S,-,2Z,a,2-,S,Z,S,-	0.66
	2Z,S,2Z,2a,-,2S,Z,3-,2S,-,2Z,a,-,Z,S,-,S,a	0.66
TOTAL (%)	8/1980	7.68

Trajectory	3	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a	(-)
MD Conf.	2Z,a,-,S,-,S,Z,-,3S,2-,S,-,Z,2S,-,+,3S,a,-	1.44
	2Z,a,-,S,-,S,Z,-,3S,2-,S,-,Z,2S,-,a,3S,a,-	1.00
	2Z,a,-,S,-,S,2-,3S,2-,S,-,Z,2S,-,+,3S,a,-	0.88
	Z,2S,-,+,3S,a,2-,Z,a,-,S,-,S,Z,-,3S,2-,S,-	0.80
	Z,2S,-,a,3S,a,2-,Z,a,-,S,-,S,Z,-,3S,2-,S,-	0.74
	2Z,a,-,S,-,S,-,Z,3S,2-,S,-,Z,2S,-,+,3S,a,-	0.64
	Z,2S,Z,a,Z,S,Z,a,-,Z,-,a,2-,2S,Z,-,3S,2-,S,-	0.58
	2Z,a,-,S,-,S,2-,3S,2-,S,-,Z,2S,-,a,3S,a,-	0.54
TOTAL (%)	8/2429	6.62

Trajectory	average	
CD26	name D-II	w (%)
Starting Conf.		
MD Conf.	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,3Z,S	1.11
	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,Z,2S,3Z,S	1.01
	4Z,S,-,2S,-,a,-,S,-,S,2-,Z,+,-,3S,2Z,-,S	0.81
	3Z,S,2Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,2Z,-,S	0.73
	2Z,S,2Z,2a,Z,2S,-,Z,2-,2S,-,2Z,a,2-,S,Z,S,a	0.65
	3Z,S,2Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,Z,S,-,S	0.63
	3Z,S,-,Z,2S,-,a,Z,-,2S,2-,Z,+,-,3S,2Z,-,S	0.55
	3Z,S,2Z,2S,-,a,2Z,2S,2-,Z,+,-,3S,Z,S,-,S	0.55
TOTAL (%)	8/5630	6.05

Figure 7.17: Conformations from CD26 water MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.2.4.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 7.18] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 7.19]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD26	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	1221				
	2	1980	5630	5630	1876.67	**0.00**
	3	2429				

Figure 7.18: *Lambda index* –trajectories overlapping ratio- for CD26 at the end of the MD calculations in water.

- The three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 7.19]. Besides, the saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 7.18], therefore, calculations are not trustworthy: on the one hand, trajectories do not converge; on the other hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturate, information about the conformational space is unreliable.
- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is underexplored.

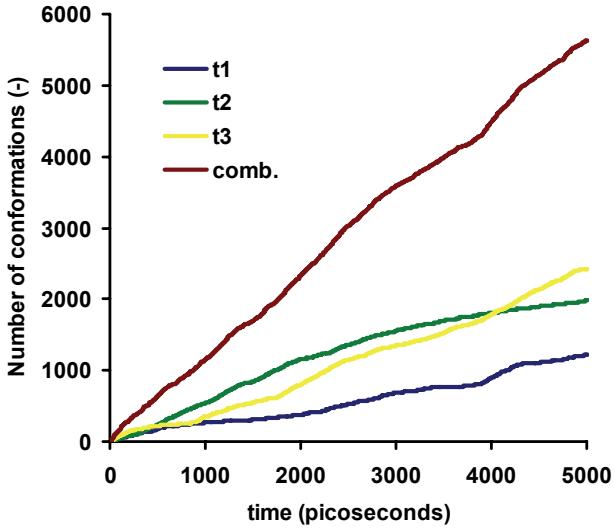


Figure 7.19: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD26 in water.

The efficiency in the individual exploration of the conformational space follows the order t3 ≈ t2 > t1. However, the total number of conformations in the combined trajectory is three times bigger, meaning that they cover separated areas. The main remarkable phenomenon in the series is the smooth continuity observed in the exploration rate of t1, t2 and t3.

The table in [Fig. 7.17] again suggests that the two-group behaviour vanishes at the expense of the *conformational noise*, which seems to be a common behaviour in large cyclodextrins. The Pareto diagram [Fig. 7.20] shows –from an optimistic point of view- that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, in fact, it is between 55% and 75%. This scenario is different from that one observed in CD14 and CD21 in water solvent and also in CD26 in gas phase; in those cases the situation was slightly better than now.

CD26		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	1221	1980	2429	5630
20% of Conf.	(n)	244	396	486	1126
Acc. Stat. Weight	(%)	**74.58**	**61.98**	**55.40**	**64.52**

Figure 7.20: Pareto. CD26 MD trajectories in solvent water.

The best 20% conformations represent between 55% and 75% of their trajectories. Similar information is shown in [Fig. 7.17]: the first 8 conformations of t1, t2, t3 and the combined trajectories were shown with their statistical weights: 18%, 8%, 7% and 6%. These ratios are not good enough to support that they can be considered as the main conformations. Then, the conformational noise is –step by step- starting to rule the general conformational behaviour.

7.2.4.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- clearly states that the three trajectories are completely unconnected.

7.2.4.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

Figure 7.21: Average conformations from CD26 water MD trajectories t1, t2 and t3.

The situation in [Fig. 7.21] is similar to CD21 calculations –in gas phase and also "in vacuo"-. D-II analyses are partially useful and they contrast with the average conformations obtained with ptraj. Molecular Dynamics trajectories tend to remain reasonably close to the starting conformations and then, D-II, being extremely sensitive to small conformational changes, reveals inadequate to describe CD26.

- Under the new criterion, the three of them; t1, t2 and t3 can be regarded as different conformations.
- Hydrogen bonds are very important in stabilizing conformations. Different types of arrangements coexist at the same time to allow loops, chains and helices because of the secondary structure.
- The directionality involving hydrogen bonds is a less important phenomenon not fully used in the family of the large CDs. However, partial sequences are detected.

7.2.5 CD28

CD28 is the last and largest macromolecule studied in this work that belongs to the group of the *Large Cyclodextrins*. Similarly to what was done in the cases of CD14, CD21 and CD26, three conformations were selected for the MD calculations in water. The set of Molecular Dynamics carried out for CD28 included three different starting conformations: the first one [**3Z,S,+,Z,S,Z,a,2Z,S,2Z,2S,a,-,Z,S,-,2Z,S,a,+,a,-**]; the second one [**5Z,a,-,Z,S,Z,S,+,Z,S,-,a,2Z,S,Z,2S,-,a,2S,-,a**] and the third one

[**2Z,S,2Z,S,-,a,Z,-,2S,Z,a,2S,-,S,Z,-,S,-,2S,Z,3S**]. All of them were randomly selected and are the same as those employed in the gas phase MD calculations.

7.2.5.1 Population Analysis

Some of the most representative results can be found in [Fig. 7.22]. As was previously done in gas phase calculations, 8 conformations have been selected, explaining between 9% and 30% of the system, which is a significant difference in comparison to the rates measured in the gas phase calculation –between 27% and 52%-. The increment in the ratios is a good result; in fact it is, surprisingly, a better result than that one measured in the series of gas phase MD calculations which is, in principle, a contradiction to the rule already mentioned in the case of CD26: *"more glucose units, more flexibility"*. Some conclusions are shown hereafter:

- A total number of 5411 conformations were found.
- The total number of conformations found in the individual trajectories t1 and t2 is almost the same –respectively 2150 and 2112-. Trajectory t3 is, however, noticeably smaller with 1149. The total number of conformations found in the combined trajectory is nearly three times bigger –as said, 5411-.
- The trajectories t1, t2, t3 do not overlap and the combined trajectory does not have in common any of the conformations.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**6Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,-,S,-,+,a**], [**2Z,S,a,2-,2Z,3-,Z,S,Z,3S,Z,-,2S,-,Z,3S,-,a**], [**Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,2-,2S,Z,S**] and [**Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,2-,2S,Z,S**]. Again, as was found in the case of CD26, their ratios are smaller than those measured for the small cyclodextrins and the gas phase calculations. But amazingly higher than the ones measured in CD26.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase, this result meaning that the solvent influence in the molecular behaviour is an important one.

- The starting conformations in t1, t2 and t3 are not detected during the MD calculations. Furthermore, the highly rated conformations –generally speaking- are so poorly weighted that they cannot be considered as important. The result suggests that, on the one hand, MD calculations do not necessarily get stuck in the starting conformations, and on the other hand, every trajectory covers areas that preferably surround certain positions.

Trajectory	1	
CD28	name D-II	w (%)
Starting Conf.	3Z,S,+,Z,S,Z,a,2Z,S,2Z,2S,a,-,Z,S,-,2Z,S,a,+,a,-	(-)
MD Conf.	6Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,-,S,-,+,a	2.72
	2Z,S,a,Z,-,S,Z,-,2S,a,-,Z,S,-,2Z,S,a,+,a,-,S,-,Z,2S	2.70
	6Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,3Z,+,a	2.46
	2Z,S,a,Z,-,S,Z,-,2S,a,2Z,S,-,2Z,S,a,+,a,-,S,-,Z,2S	1.04
	3Z,S,-,2S,a,-,Z,S,-,2Z,S,a,+,a,-,S,-,Z,2S,3Z,a	1.00
	3Z,S,-,Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,-,S,-,+,a	0.80
	3Z,S,2Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,-,S,-,+,a	0.72
	3Z,S,-,2S,a,-,Z,S,-,2Z,S,a,+,a,-,S,-,Z,2S,2Z,S,a	0.64
TOTAL (%)	8/2150	12.08

Trajectory	2	
CD28	name D-II	w (%)
Starting Conf.	5Z,a,-,Z,S,Z,S,+,Z,S,-,a,2Z,S,Z,2S,-,a,2S,-,a	(-)
MD Conf.	2Z,S,a,2-,2Z,3-,Z,S,Z,3S,Z,-,2S,-,Z,3S,-,a	2.44
	Z,3S,-,Z,a,2-,Z,3S,-,a,2S,-,a,-,S,5-,Z,S	1.30
	2Z,3-,Z,S,-,3S,Z,-,2S,-,Z,3S,-,a,Z,S,Z,a,2-	1.00
	2Z,S,a,2-,2Z,3-,Z,S,Z,3S,2-,2S,-,Z,3S,-,a	0.80
	2Z,3-,Z,S,Z,3S,Z,-,2S,-,Z,3S,-,a,Z,S,Z,a,2-	0.80
	2Z,S,a,2-,2Z,3-,Z,S,Z,3S,2-,2S,-,S,Z,2S,-,a	0.78
	2Z,S,a,2-,2Z,3-,Z,S,-,3S,Z,-,2S,-,Z,3S,-,a	0.78
	Z,3S,2-,a,2-,Z,3S,-,a,2S,-,a,-,S,5-,Z,S	0.74
TOTAL (%)	8/2112	8.64

Trajectory	3	
CD28	name D-II	w (%)
Starting Conf.	2Z,S,2Z,S,-,a,Z,-,2S,Z,a,2S,-,S,Z,-,S,-,2S,Z,3S	(-)
MD Conf.	Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,2-,2S,Z,S	9.10
	Z,2S,Z,S,Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,-	8.72
	Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,-,S,-,2S,Z,S	3.34
	Z,2S,-,4S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,Z,S,Z,-,S	2.12
	Z,2S,-,S,Z,-,S,-,a,2-,Z,S,Z,a,2S,-,2S,Z,2-,2S,Z,S	1.88
	Z,2S,Z,S,Z,2S,-,S,Z,-,S,-,a,2-,Z,S,Z,a,2S,-,2S,Z,-	1.78
	Z,2S,Z,2-,2S,Z,S,Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S	1.48
	Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,-,Z,-,2S,Z,S	1.46
TOTAL (%)	8/1149	29.88

Trajectory	average	
CD28	name D-II	w (%)
Starting Conf.		
MD Conf.	Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,2-,2S,Z,S	3.03
	Z,2S,Z,S,Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,Z,-	2.91
	Z,2S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,-,2S,-,S,-,2S,Z,S	1.11
	6Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,-,S,-,+,a	0.91
	2Z,S,a,Z,-,S,Z,-,2S,a,-,Z,S,-,2Z,S,a,+,a,-,S,-,Z,2S	0.90
	6Z,S,a,+,a,-,S,-,2Z,S,2Z,S,a,-,Z,S,3Z,+,a	0.82
	2Z,S,a,2-,2Z,3-,Z,S,Z,3S,Z,-,2S,-,Z,3S,-,a	0.81
	Z,2S,-,4S,-,S,Z,-,S,-,a,2-,2S,Z,a,2S,Z,S,Z,-,S	0.71
TOTAL (%)	8/5411	11.20

Figure 7.22: Conformations from CD28 water MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.2.5.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 7.23] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 7.24]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD28	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	2150	5411	5411	1803.67	**0.00**
	2	2112				
	3	1149				

Figure 7.23: *Lambda index* –trajectories overlapping ratio- for CD28 at the end of the MD calculations in water.

- All three trajectories; t1, t2 and t3, evolve in time towards better explorations of the conformational space, although none of them saturates [Fig. 7.24]. The saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 7.23]. Therefore, calculations are not trustworthy because; on the one hand, trajectories do not converge; on the other hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since

none of them saturates, information about the conformational space is unreliable.

- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces. Then, the system is underexplored.

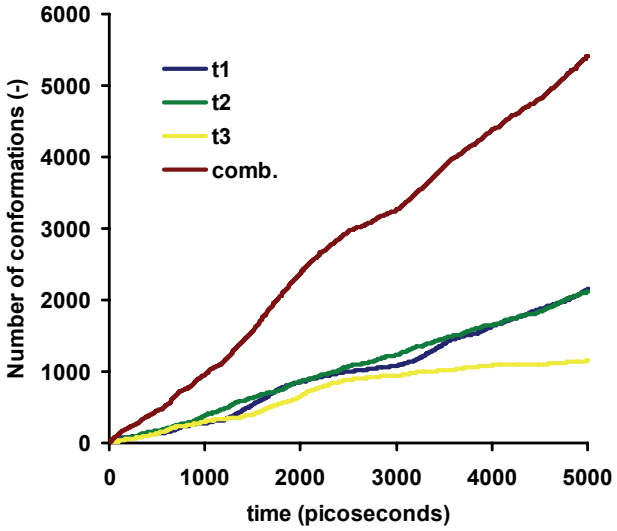


Figure 7.24: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD28 in water.

The efficiency in the individual exploration of conformational space follows the order t1 ≈ t2 > t3, although, under 3000 picoseconds the three trajectories evolve in a similar way. The total number of conformations in the combined trajectory is –as usual in the *large cyclodextrins* group- three times bigger, meaning that they cover separated areas. The quasi-linear behaviour in the series of the individual trajectories –that meet at zero steps- and on the other hand, the smooth continuity observed in the exploration rate of t1, t2 and t3 –where t1 and t2 grow almost in parallel- are remarkable points.

Examining the table in [Fig. 7.22] it is found that the two-group behaviour vanishes at the expense of the *conformational noise*. The Pareto diagram [Fig. 7.25] seems to suggest again –from an optimistic point of view- that the trajectories could still be divided into the two groups, although the average ratio explained by the best 20% conformations is under 80%, in fact, it is between 60% and 76%. This situation is –in average- similar to that one observed in CD26 in water solvent, but slightly underrated in comparison with CD28 in gas phase.

CD28		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	2150	2112	1149	5411
20% of Conf.	(n)	430	422	230	1082
Acc. Stat. Weight	(%)	**61.42**	**60.30**	**75.70**	**66.23**

Figure 7.25: Pareto. CD28 MD trajectories in solvent water.

The best 20% conformations represent between 60% and 76% of their trajectories. Similar information is shown in [Fig. 7.22]: the first 8 conformations of t1, t2, t3 and the combined trajectories were shown with their statistical weights: 12%, 9%, 30% and 11. These ratios –although slightly better than those in CD26- are not good enough to support that they can be considered as the main conformations. Then, it may be said that the conformational noise is ruling the general conformational behaviour; which turns out to be a general property of large cyclodextrins.

7.2.5.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- clearly states that the three trajectories are completely unconnected.

7.2.5.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

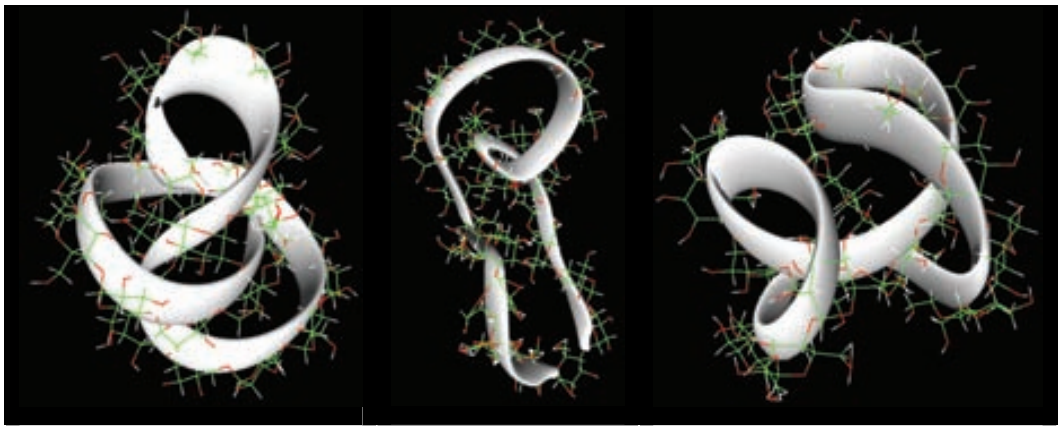


Figure 7.26: Average conformations from CD28 water MD trajectories t1, t2 and t3.

We are facing the same scenario found in CD21 and CD26. The images of the average structures [Fig. 7.26] contradict once more the results obtained from the descriptor analysis concluding that D-II is extremely sensitive to small conformational changes and therefore inadequate to describe CD28. Furthermore, Molecular Dynamics trajectories tend to remain close to the starting conformations.

- Under the new criterion, the three of them; t1, t2 and t3 can be regarded as different conformations. CD28 folds creating some loops although not clearly seen because of the entangled conformations –in t1 and t3-. Meanwhile, a loose and partially extended conformation with only a small loop is obtained –in t2-.
- Hydrogen bonds are very important to stabilise conformations and different types of arrangements coexist at the same time to allow the secondary structure.
- The directionality involving hydrogen bonds is a less important phenomenon not fully used in the family of the large CDs. However, partial sequences are detected.

7.3 Molecular Dynamics in Benzene Solvent (C_6H_6)

As said, the point of this calculation is comparing the results for benzene with those previously obtained in gas phase and water to evaluate the influence of solvent polarity in the conformational space of cyclodextrins.

Similarly to the case of the water calculations, the system is conceived as a truncated octahedral box where the cyclodextrin is placed in the middle, and the solvent –the explicit benzene solvent model half-developed in our group- surrounds the molecule until it fills up the remaining space within the unit cell. The series were carried out following the schema in [Fig. 7.27] and the full calculation includes four steps.

- Step zero is a restricted molecular energy minimization. It is designed to remove –if necessary- any remaining residual tension generated during the solvation process. By default algorithms –the first ten steps of *steepest descent* followed by *conjugated gradient*- are employed, including up to 50,000 steps or by-default termination energy criterion.

- The first step is the *heating slope* and *equilibration* process under NVT conditions –*canonical ensemble*-: 300 picoseconds length, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, 298 K in the equilibrium and external constraints to avoid chair-to-boat conformational interchanges in glucoses. Coupling constants are modulated along the simulation to avoid *blow-up* terminations.
- The second step is the *equilibration* process under NPT conditions –*isothermal-isobaric ensemble*-: 300 picoseconds length, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, bath thermal coupling of 0.5 picoseconds, temperature 298 K in the equilibrium and pressure 1 atmosphere. External constraints are removed and coupling constants are kept unchanged along the simulation.
- The third step is the *sampling* process under NPT conditions –*isothermal-isobaric ensemble*-. Its length is 5,000 picoseconds, timestep of 1 femtosecond, cutoff of 12 angstrom, PBC, 298 K, pressure 1 atmosphere, bath thermal coupling of 0.5 picoseconds and sampling frequency of 1 snapshot per picosecond –the trajectory is stored in this step for further analysis-. External constraints –to avoid chair-to-boat conformational interchanges in glucoses- are removed and coupling constants are kept unchanged along the simulation.

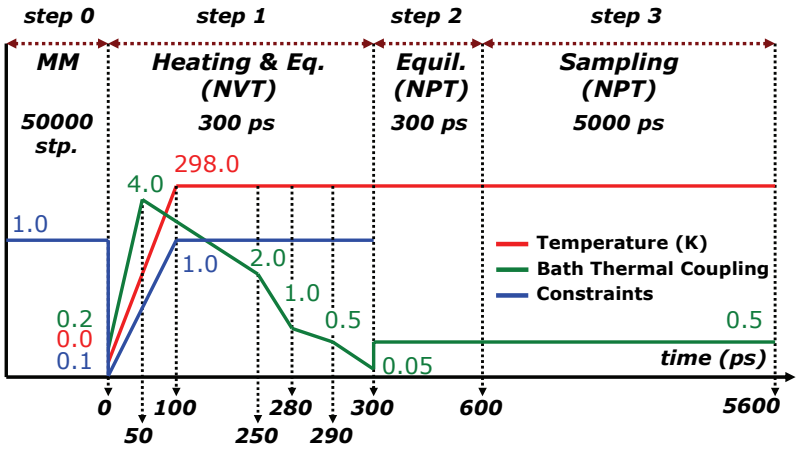


Figure 7.27: Schematic depiction of the 4-step benzene-solvent Molecular Dynamics protocol. Coloured lines, although not scaled, represent the dynamically coupled control parameters.

Detailed information regarding explicit Benzene solvent model MD conditions and further computational aspects can be found in the DVD: annex 10.1.1.6 and its sub-annexes (chapter X AMBER 7 files).

7.3.1 CD26

As was said, CD26 is another one of the four cyclodextrins –alongside with CD14, CD21 and CD28- that belongs to the group of the *Large Cyclodextrins*. The new series of MD calculations in benzene solvent were planned as a way of evaluating the effect of solvent polarity in the cyclodextrins. Having this in mind, CD26 was selected and then, three different series of three calculations were done; firstly, "in vacuo" (gas phase, with dipolar moment of 0 Debyes); secondly, in water (polar solvent with a dipolar moment of about 1.85 Debyes); and finally, in benzene (non-polar solvent with dipolar moment of 0 Debyes).

Therefore, the set of Molecular Dynamics carried out in benzene solvent included three different starting conformations: the first one [**3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S**]; the second one [**2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S**] and the third one [**3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a**]. All of them were randomly selected and are the same as those employed in the gas phase and water solvent MD calculations.

7.3.1.1 Population Analysis

The most representative results of the population analysis can be found in [Fig. 7.28]. As was done in the case of gas phase and water solvent calculations, 8 conformations have been selected. They explain between 21% and 53% of the system, which is a significant difference in comparison with the rates measured in the gas phase –between 20% and 37%- and especially in the water solvent –between 6% and 18%- MD calculations. The noticeable increase in the ratios is an interesting result –in fact better than those ones in the series of gas phase and water MD calculations- meaning that, somehow, the influence of the solvent is important not only in a quantitative, but also in a qualitative way. Some conclusions are shown hereafter:

- A total number of 1806 conformations were found –considerably less than the number obtained, 5630, in the water calculations-.

- The total number of conformations found in the individual trajectories t1, t2 and t3 is unquestionably different –respectively 311, 568 and 927-, being the total number of conformations found in the combined trajectory nearly three times bigger –about 1806-. These figures are again considerably smaller –about one third- than those obtained in the series of homologous water MD calculations.
- The trajectories t1, t2, t3 and the combined trajectory do not have in common any of the conformations. In fact, none of them overlaps.
- The most highly rated conformations in t1, t2, t3 and the combined trajectory are: [**4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,2S,-,S**], [**3Z,S,a,-,Z,3S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-**], [**2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a,Z,-,Z,a,-,Z,S**] and [**4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,2S,-,S**]. Although their ratios are lower than those measured for the small cyclodextrins, they are –generally speaking- higher than the ratios obtained from the calculations in gas phase and water solvent.
- The most populated conformations detected by MD in the combined trajectory are not coincident with the highly populated ones found in the SA Conformational Search and the MD in gas phase and water solvent, this result meaning that the solvent influence in the molecular behaviour is an important one.
- The starting conformations in t1, t2 and t3 are not detected during the MD calculations. Furthermore, although the highly rated conformations are better weighted that those ones in the gas phase and water solvent calculations, still they cannot be considered as important. The results suggest that; firstly, MD calculations do not necessarily get stuck in the starting conformations; secondly, every trajectory covers areas that preferably surround certain positions; and finally, the areas explored are strongly influenced by the solvent.

Trajectory	1	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,3Z,S,2Z,2S,-,a,-,Z,2S,2-,2Z,2S,Z,S	(-)
MD Conf.	4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,2S,-,S	23.24
	4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,S,Z,-,S	6.60
	4Z,S,2Z,2S,-,a,2Z,2S,2-,Z,3S,Z,2S,-,S	6.14
	2Z,2S,-,a,-,Z,2S,2-,Z,3S,-,S,Z,-,2S,-,2Z,S	4.32
	4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,-,2S,-,S	3.42
	2Z,2S,-,a,-,Z,2S,2-,Z,3S,-,2S,-,2S,-,2Z,S	3.36
	3Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,2S,-,2S	2.98
	3Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,S,Z,-,2S	2.72
TOTAL (%)	8/311	52.78

Trajectory	2	
CD26	name D-II	w (%)
Starting Conf.	2Z,2S,Z,-,Z,2S,-,2Z,a,-,Z,S,-,S,a,2Z,S,Z,S,a,S	(-)
MD Conf.	3Z,S,a,-,Z,3S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	7.18
	3Z,S,a,-,Z,S,Z,S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	6.42
	3Z,S,a,2Z,3S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	6.08
	3Z,S,a,2Z,S,Z,S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	5.04
	3Z,S,a,-,Z,S,Z,S,a,Z,S,Z,2S,2-,Z,2S,2-,S,a,-	4.24
	3Z,S,a,2Z,S,Z,S,a,Z,S,Z,2S,2-,Z,2S,2-,S,a,-	3.96
	3Z,S,a,-,Z,S,Z,S,a,Z,S,Z,2S,Z,-,Z,2S,2-,S,a,-	3.88
	3Z,S,a,2Z,S,Z,S,a,Z,S,Z,2S,Z,-,Z,2S,2-,S,a,-	3.82
TOTAL (%)	8/568	40.62

Trajectory	3	
CD26	name D-II	w (%)
Starting Conf.	3Z,S,-,Z,2S,2Z,-,S,-,Z,2S,-,a,3S,a,-,2Z,a	(-)
MD Conf.	2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a,Z,-,Z,a,-,Z,S	5.14
	2Z,-,3S,2-,Z,-,2S,Z,S,a,Z,S,Z,a,Z,-,Z,a,-,Z,S	3.30
	3Z,a,-,Z,S,2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a	2.92
	2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a,2Z,-,a,-,Z,S	2.84
	3Z,a,-,Z,S,2Z,-,3S,2-,Z,-,Z,S,Z,S,a,2S,Z,a	1.90
	2Z,S,2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a,2Z,-,a	1.72
	5Z,-,3S,2-,S,-,Z,2S,Z,a,Z,S,Z,a,Z,-,Z,a	1.46
	2Z,S,2Z,-,3S,2-,Z,-,2S,Z,S,a,Z,S,Z,a,Z,-,Z,a	1.46
TOTAL (%)	8/927	20.74

Trajectory	average	
CD26	name D-II	w (%)
Starting Conf.		
MD Conf.	4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,2S,-,S	7.75
	3Z,S,a,-,Z,3S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	2.39
	4Z,S,2Z,2S,-,a,-,Z,2S,2-,Z,3S,Z,S,Z,-,S	2.20
	3Z,S,a,-,Z,S,Z,S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	2.14
	4Z,S,2Z,2S,-,a,2Z,2S,2-,Z,3S,Z,2S,-,S	2.05
	3Z,S,a,2Z,3S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	2.03
	2Z,-,3S,2-,Z,-,2S,Z,S,a,2S,Z,a,Z,-,Z,a,-,Z,S	1.71
	3Z,S,a,2Z,S,Z,S,a,S,Z,3S,Z,-,Z,2S,2-,S,a,-	1.68
TOTAL (%)	8/1806	21.95

Figure 7.28: Conformations from CD26 benzene MD trajectories 1, 2, 3, average trajectory and their statistical weights. Names emphasized in grey help identifying MD dependence on starting conformation.

7.3.1.2 Overlapping Analysis I: Accumulative Lambda Index

The *lambda index* is measured at the end of the MD calculations [Fig. 7.29] from the total number of conformations in t1, t2, t3 and the combined trajectory [Fig. 7.30]. Likewise, the *dynamical lambda index* calculated during all the calculations was always valued 0%.

CD26	i	^{i}n	t	$q = \Sigma^{i}n$	<n>	λ(%)
MD	1	311				
	2	568	1806	1806	602.00	**0.00**
	3	927				

Figure 7.29: *Lambda index* –trajectories overlapping ratio- for CD26 at the end of the MD calculations in benzene.

- The three trajectories; t1, t2 and t3, evolve in time towards better explorations of conformational space, although none of them saturates [Fig. 7.30]. Besides, the saturation plot of lambda index vs. time was always 0% and the value of the lambda index at the end of the calculation was 0% [Fig. 7.29], therefore, calculations are –as usual in the large cyclodextrins- not trustworthy because; on the one hand, trajectories do not converge; on the other hand, the ratios are really low.
- The total number of conformations in the combined trajectory and the lambda index do not reach saturation. Nevertheless, they are still good parameters for monitoring the evolution of the system.
- The three individual trajectories t1, t2, and t3 cannot separately explain the whole system. The combined trajectory is a better approach. However, since none of them saturates, information about the conformational space is unreliable.
- Trajectories t1, t2, and t3 have in common 0% of their conformational spaces; therefore, the system is underexplored.

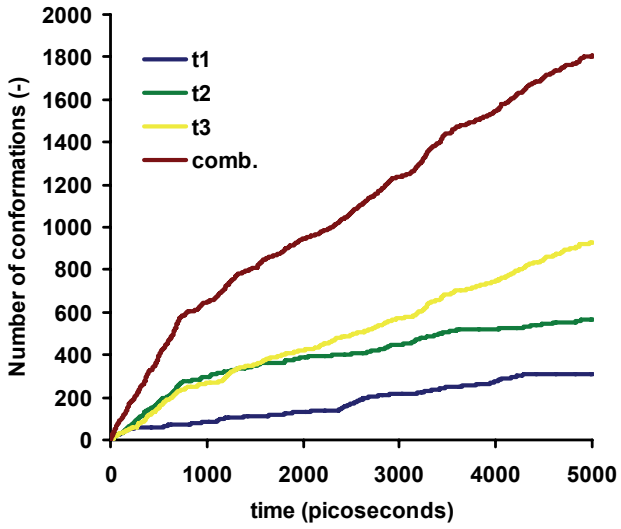


Figure 7.30: Number of conformations in *t1, t2, t3* and total number of conformations in the combined trajectory, *comb* for CD28 in benzene.

The efficiency in the individual exploration of conformational space follows the order t3 > t2 > t1, being the total number of conformations in the combined trajectory about three times bigger, meaning that they cover separated areas. The main remarkable phenomena in the series are; on the one hand, the smooth continuity observed in the

exploration rate of t1, t2 and t3; on the other hand, the noticeable decrease in the number of conformations in the individual and combined trajectories when compared with the equivalent MD series of water solvent calculations. While the combined trajectory in water found about 5630 conformations, now the number has been reduced to 1806. Since the only difference is the solvent, it can be said that the solvent effect is determinant in the conformational properties of the system: Non-polar solvents –benzene in our case- seem to encourage folding towards tighter, closer and more compact conformations favouring intramolecular interactions rather than solvent-solute ones.

The Pareto diagram in [Fig. 7.31] is absolutely amazing. The rates for the most important conformations have increased considerably in comparison with those obtained for the CD26 water solvent simulations, which means that the two-group behaviour –including the *main conformations* and the *conformational noise*- is again detected. Now, the average ratio explained by the best 20% conformations is over 84%, in fact, it is between 76% and 90%. This scenario is noticeably better that the one observed in the CD26 equivalent calculations in gas phase and water solvent. The new situation is, undoubtedly, the consequence of the solvent influence.

CD26		Pareto			
		t1	t2	t3	comb.
Number of Conf.	(n)	311	568	927	1806
20% of Conf.	(n)	62	114	185	361
Acc. Stat. Weight	**(%)**	**90.24**	**85.72**	**76.04**	**84.18**

Figure 7.31: Pareto. CD28 MD trajectories in solvent benzene.

As was said, the best 20% conformations represent between 76% and 90% of their trajectories. Similar information is shown in [Fig. 7.28]: the first 8 conformations of t1, t2, t3 and the combined trajectories were shown with their statistical weights: 53%, 41%, 21% and 22%. Apparently, these ratios are very good –in comparison to those obtained in water solvent; 18%, 8%, 7% and 6%- however, the situation is different. The current figures are not the consequence of an effective conformational search, but the expected behaviour of a molecule which is not soluble in non-polar solvents. Under these conditions, the macroring tends to fold into collapsed conformations that hidden polar groups inside the pockets and offer non-polar parts to the solvent. This makes the molecule to clot into a small number of conformations.

7.3.1.3 Overlapping Analysis II: Omega Index

The table is omitted because the omega index was 0% for all the conformations found. The null overlapping ratio between trajectories –the lambda index value is 0%- clearly states that the three trajectories are completely unconnected.

7.3.1.4 Average Structures

Protein-like visualizations of the type strands, ribbons and bands are employed:

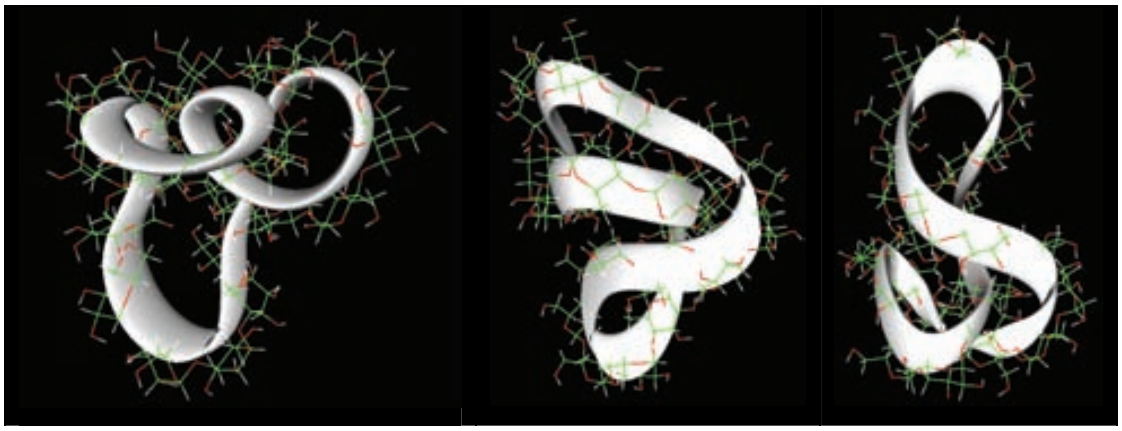


Figure 7.32: Average conformations from CD26 benzene MD trajectories t1, t2 and t3.

Although a non-polar solvent has been used in these calculations –benzene-, the results in general [Fig. 7.32] remain almost identical to those for the polar solvent –water-. D-II results contradict the images of the average structures. The reason has been already explained; Molecular Dynamics trajectories tend to remain reasonably close to the starting conformations and then, D-II, being extremely sensitive to small conformational changes, reveals inadequate to describe CD26.

- Examining the images, the three of them –t1, t2 and t3- can be considered as different conformations.
- Hydrogen bonds are very important stabilizing conformations, where different types of arrangements coexist at the same time to help secondary structure (loops, helices, straps).

- The directionality involving hydrogen bonds is a less important phenomenon not fully used in the family of the large CDs. However, partial sequences are detected.

7.4 GROUP PROPERTIES II: The Effect of Solvent in CD26

Up to this point the properties and results of all the cyclodextrins have been –more or less- individually explained. However, sporadic comparisons involving homologous calculations were included in several sections.
Now, a comparative study in CD26 is discussed within the current section involving the effect of the solvent polarity in the Molecular Dynamics calculations. With the purpose of evaluating the influence of the presence/absence of solvent and the polarity in MD calculations, the series of MD were carried out under different environments.

7.4.1 Brief summary of solvent-solute interactions

Molecular behaviour in solution is complex (Further details involving solvation models in computational chemistry can be consulted in chapter IX) and many different interactions drive the global dynamics. In this sense, the effects can be divided for our purposes in two parts:

- In the first place, the contributions of the *non-bonded interactions*: hydrogen bonds, electrostatic in general, van der Waals, etcetera. This is usually the main point and also the one that computational chemists mostly care of when building their molecules.
- In the second place, the contributions coming from the *pure mechanical motion* originated by the thermal energy and represented by colliding molecules. This point is commonly regarded as a secondary topic although it models an important part of the solvent-solute interactions, unable to be explained by means of continuum solvent models. In fact, solute molecules have to *hit* solvent molecules to make room prior to changing their conformations or just make their way and move through the solvent in the diffusion phenomenon.

This study tries to evaluate the solvent effect in two ways:

1) The effect of polarity when running MD calculations in polar solvent –water- and non-polar solvent –benzene-.
2) The effect of the presence/absence of solvent when running MD calculations. A practical note: since gas phase Molecular Dynamics can be considered as MD calculations using a non-polar solvent with no mass, this point should be compared with the homologous benzene calculation to measure the effect.

7.4.2 Comparative Molecular Dynamics

The three series of three Molecular Dynamics –in gas phase, water and benzene- previously explained have been jointly analysed in this section. The experimental conditions were appropriately explained [*chapter VI, section 6.2.8; and current chapter, sections 7.2.4; and 7.3.1*]. The whole set of analysis were made on the grounds of several data: D-II, solvation energy, hydrogen bonds occupancy ratio and average structures.

7.4.2.1 Total Number of Conformations according to D-II

The graphics in [Fig. 7.33] represent the total number of conformations in t1, t2, t3 and [Fig. 7.34] the combined trajectory according to D-II. Every plot shows the same MD calculation under three different conditions –"in vacuo", in water and in benzene-.

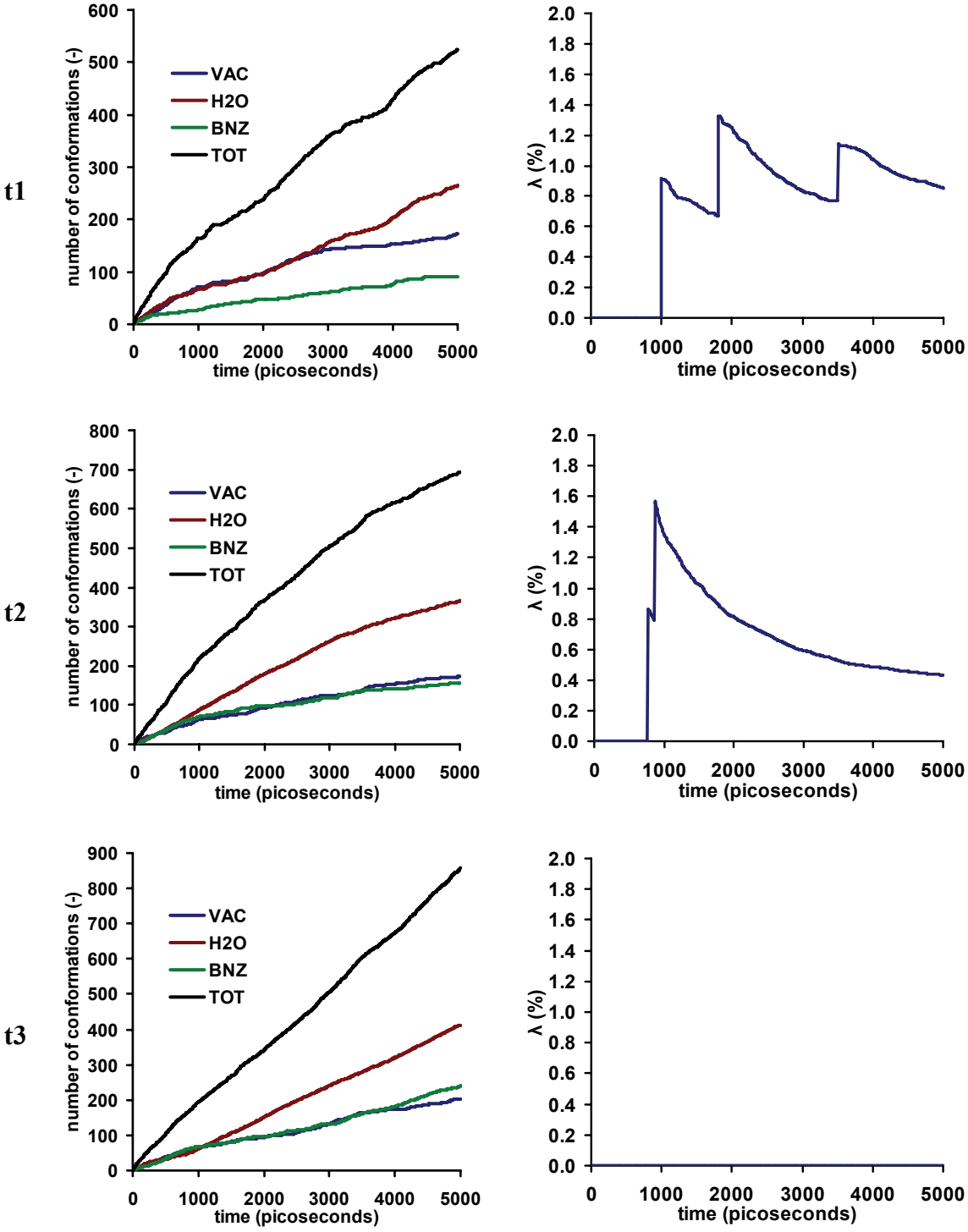


Figure 7.33: Left column: Number of conformations in gas phase, in water and also in benzene in t1, t2, t3 and combined trajectory for CD26. In this plot, the *combined trajectory* –in black- represents the combination of the gas phase trajectory + the water trajectory + the benzene trajectory for t1; and similarly for t2 and t3. Right column: Dynamic Lambda indexes for the trajectories in gas phase, water and benzene: Lambda index for t1 –in "in vacuo", water and benzene trajectories- is calculated. Identical procedure is done on t2 and t3 groups of trajectories.

Since the original production times and sampling frequencies were different, a data pre-treatment previous to the MD trajectories analysis was necessary to ensure full consistency: Gas phase calculations were 10 nanoseconds long and sampling frequency was 0.1 structures per picosecond. Meanwhile, solvent calculations –both in water and in benzene- were 5 nanoseconds long, and a sampling frequency of 1 structure per picosecond was used. The compromise solution was cutting down the total simulation time in gas phase calculations to 5 nanoseconds, and the data in solvent calculations were sampled again employing the value 0.1 structures per picosecond.

The analysis of the four graphics gives some clues about the solvent effect. The MD calculations in water –polar solvent- find more conformations than those in benzene –non-polar solvent- and gas phase. This probably means that conformations in water are: flexible, less compacted, less tightly folded and more open to water, enabling favourable hydrogen bonding interactions. The situation in benzene calculations is a little different, finding fewer conformations and folding into compact and more rigid structures that hide the polar parts of the molecule in the inner side to prevent unstable solvent interactions.

When non-polar benzene and gas phase calculations are compared they seem to suggest that the behaviour is almost similar. However, it can be said that the scenario is a little better –regarding conformational exploration- when the MD simulations are performed in gas phase [Fig. 7.34]. In the end, although gas phase and benzene are "non-polar" environments, the presence of solvent molecules represents a steric hindrance just on the grounds of the mechanical inertia. This fact could be responsible for the slight decrease in the total number of conformations in the combined trajectory.

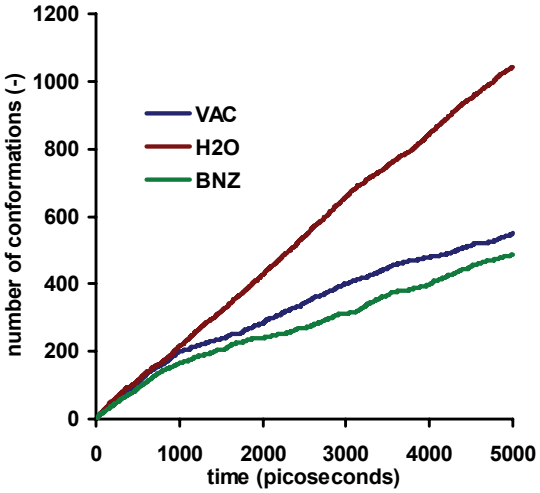


Figure 7.34: Total Number of conformations in gas phase, in water and also in benzene in the combined trajectory for CD26. In this plot, the *combined trajectory* is the combination of the t1 + t2 + t3 in gas phase; and similarly for water and benzene –the ordinary combination employed along this thesis-.

The extremely low ratios [Fig. 7.33 right column] in lambda indexes –in the case of t3 is exactly 0.0%- mean that, though the sets of trajectories t1, t2 and t3 start from the same conformations, they apparently evolve in different directions.

This is again the natural consequence of D-II extreme sensibility. This point has two readings that are worth to mention. On the one hand, D-II is perfectly valid as an objective comparison criterion because –mathematically speaking- it equally treats all the molecules and it can also be quantitatively measured. Therefore, comparisons between gas phase, water and benzene calculations can be considered trustworthy.

On the other hand, if we were questioned about the similarity of the set of conformations within a certain MD trajectory –say t1 in water, or t3 in benzene- the only answer is that the average conformation represents the main tendency, which in principle contrasts with the results from D-II –all conformations are different under this point of view-.

Then, when the dynamic lambda index states that the overlapping ratio between gas phase, water and benzene calculations for a certain trajectory is almost 0.0%, the correct reading is; *"yes, they are different, but the difference is not significant because the descriptor is extremely sensitive to small conformational changes. Had it been used a more suitable descriptor; the overlapping ratio would have been higher and also closer*

to reality". The key is again the folding property and the fact that large CDs have secondary structure.

7.4.2.2 Solvation Energy

It was said in the previous section that conformations found in water calculations featured polar groups exposed to the solvent according to favourable solvent-solute interactions. However, conformations in the benzene solvent hide their polar groups buried into the inner side.

In order to check this statement the solvation free energy of the nine trajectories was evaluated [Fig. 7.35] by means of the MM-PBSA methodology (see chapter IX methodology, in the appendix).

CD26	t1		t2		t3	
GBSOL	average	Std.Dev.	average	Std.Dev.	average	Std.Dev.
Benzene	-92.24	5.92	-109.68	5.08	-98.28	5.59
Gas Phase	-70.02	3.67	-69.99	2.96	-66.86	2.53
Water	-233.17	13.77	-248.35	16.36	-236.16	15.65

Figure 7.35: Solvation free energy for t1, t2 and t3 in benzene, gas phase and water for CD26. The relative dielectric constants employed were 80 for water, 2.27 for benzene and 0.0 for gas phase. The radius of probe was 1.4 for water, 2.8 for benzene and 1.4 for gas phase.

As can be seen, the energetic values prove that MD calculations in water were particularly favoured. The interactions are almost three times bigger in water than in benzene or gas phase, which agrees with the fact that CD26 is soluble in water.

These results seem to support the fact that MD calculations were more effective in searching conformations when polar solvents were used because favourable interactions allowed extended and more flexible conformations.

7.4.2.3 Hydrogen Bonds Occupancy Ratio

It is also possible to evaluate the hydrogen bonds by means of ptraj (appendix 10.2.1.2 in DVD). This process is done strictly under structural criteria –distances and angles compatible with viable hydrogen bonds- and therefore no energetic information is given here.

The output of the analysis is supplied in the form of occupancy ratio, which evaluates how many times a given hydroxyl group is taking part in any hydrogen bond on the basis of distances and angles. The result is then normalised to the total number of conformations of the trajectory.

This process has been applied to the entire set of hydroxyl groups in CD26 and therefore the summations of the normalised occupancy ratios per glucose range from 0 to 300.

Hbond/Glu		Total	Intramolecular	Intermolecular
benzene (non-polar solvent)	t1	167.18	167.18	0.00
	t2	158.93	158.93	0.00
	t3	157.63	157.63	0.00
gas phase (no solvent)	t1	175.13	175.13	0.00
	t2	175.69	175.69	0.00
	t3	169.92	169.92	0.00
water (polar solvent)	t1	150.34	125.67	24.67
	t2	168.43	119.57	48.85
	t3	167.89	122.86	45.03

Figure 7.36: Hydrogen bonds occupancy ratio per glucose for t1, t2 and t3 under different conditions: benzene, gas phase and water for CD26. In the case of cyclodextrins, the occupancy ratio per glucose ranges from 0 (the worst) to 300 (the best).

The table in [Fig. 7.36] seems to point to the gas phase calculations as those having the higher occupancy ratios, although the difference is not so clearly established.

Anyway, the most important phenomenon detected is that of the intermolecular interactions. While MD calculations in benzene and gas phase involved only intramolecular hydrogen bonds, the situation was absolutely different in the case of the water simulations, where both types –intramolecular and intermolecular- were detected.

Hbond/Glu (%)		Total	Intramolecular	Intermolecular
benzene (non-polar solvent)	t1	100.00	100.00	0.00
	t2	100.00	100.00	0.00
	t3	100.00	100.00	0.00
gas phase (no solvent)	t1	100.00	100.00	0.00
	t2	100.00	100.00	0.00
	t3	100.00	100.00	0.00
water (polar solvent)	t1	100.00	83.59	16.41
	t2	100.00	70.99	29.01
	t3	100.00	73.18	26.82

Figure 7.37: Hydrogen bonds percentage occupancy ratio per glucose for t1, t2 and t3 under different conditions: benzene, gas phase and water for CD26. In this case, the occupancy ratio per glucose ranges from 0% (the worst) to 100% (the best).

Quantitatively speaking, it can be seen [Fig. 7.37] that MD water calculations involve between 16% and 29% of intermolecular hydrogen bonds –those representing solvent-solute interactions- and between 71% and 84% of intramolecular hydrogen bonds.

These results confirm that conformations found in water calculations display polar groups pointing to the solvent according to favourable solvent-solute interactions, while those obtained in benzene solvent hide their polar groups into the inner side.

7.4.2.4 Average Structures

The table in [Fig. 7.38] gathers together the average structures already presented in their respective sections: gas phase, water and benzene MD calculations. Detailed information regarding the average process can be consulted there and also in the DVD: appendix 10.3.1.2.

The table shows the macromolecular overall behaviour as a consequence of the solvent influence in the molecular dynamics by means of depicting average tendencies.

Several conclusions can be derived from the images:

- Viable molecules are obtained, meaning that the individual conformations within every trajectory are not very different.
- The images from the average structures contradict the results obtained by D-II: apparently, D-II is not the appropriate descriptor when folding is present.
- Molecular Dynamics tend to remain close to the starting conformations: In this sense, the average conformation in t1 seems to be particularly stable since it was detected in the gas phase calculations, in the water calculations and also in the benzene calculations. A similar situation was detected in t3, where the average conformations in benzene and water are almost the same, and the one in gas phase could be similar if it collapsed into a tightened conformation.
- The situation in t2 is different. Anyway, it can be seen than the average conformations in benzene and water are quite similar: just rotating the water conformation 90 degrees counterclockwise and then slightly folding the

extremes the conformation in benzene is almost obtained. The average conformation "in vacuo" is different although it can be derived from the benzene conformation by increasing the loop in the central part.

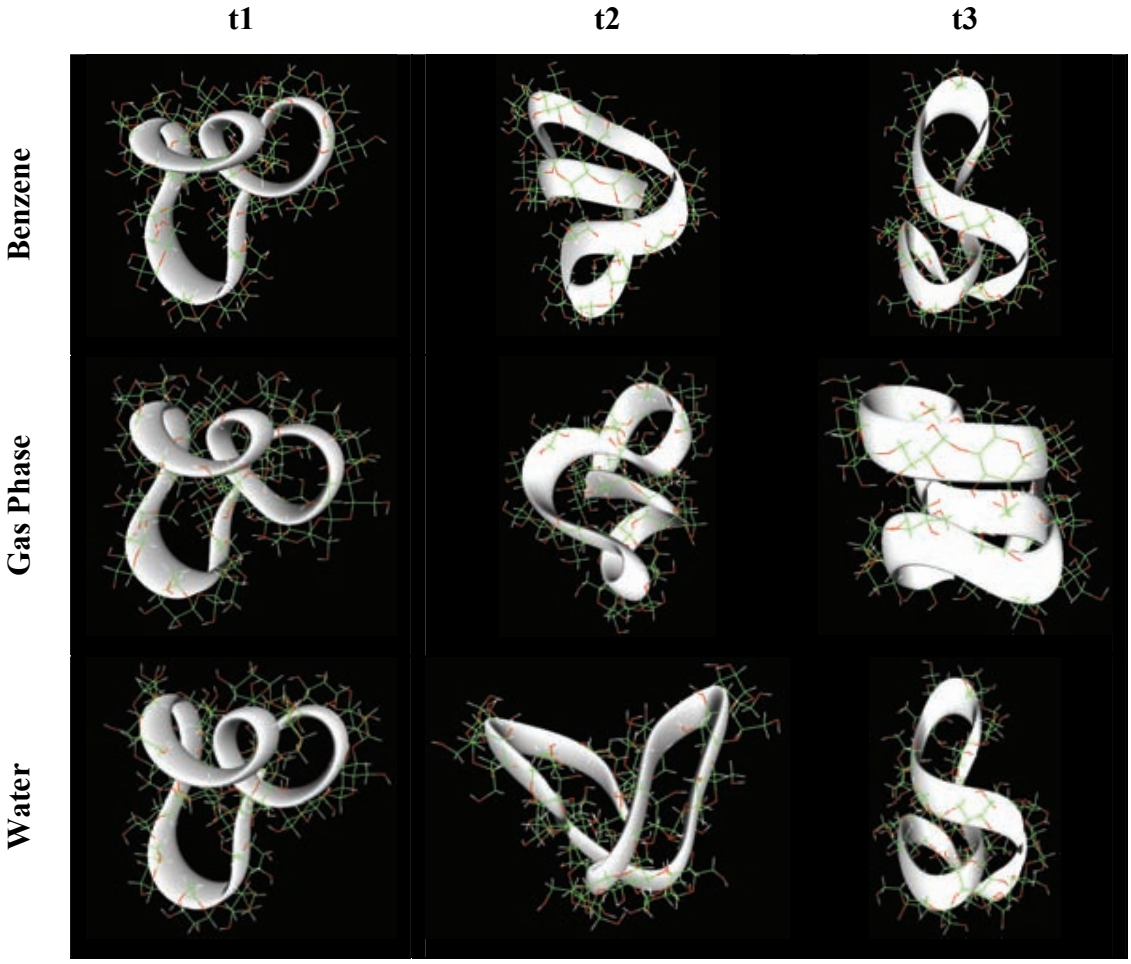


Figure 7.38: CD26 average structures of t1, t2 and t3 under different experimental conditions: benzene, gas phase and water.

7.4.3 Results

- The solvent effects can be measured qualitatively and quantitatively.
- Polar solvents favour better explorations of conformational space in cyclodextrins. This point can be easily proved by examining the analysis of the total number of conformations, the solvation free energies and the hydrogen bonds.
- Under similar dielectric constants, calculations in gas phase explore the conformational space slightly better than those in real solvent: The molecule in

gas phase does not need to collide against solvent molecules to change conformations or displace.

- Molecular Dynamics calculations tend to remain close to the starting conformations and cover areas in the vicinity. Long simulation times are necessary to overcome the handicap inherent to the methodology.
- Descriptor II is inadequate to model large cyclodextrins. It is extraordinary sensible to small conformational changes and therefore is unable to achieve a reasonable level of generality.

7.5 SUMMARY OF RESULTS

This section contains the most important results described in this chapter:

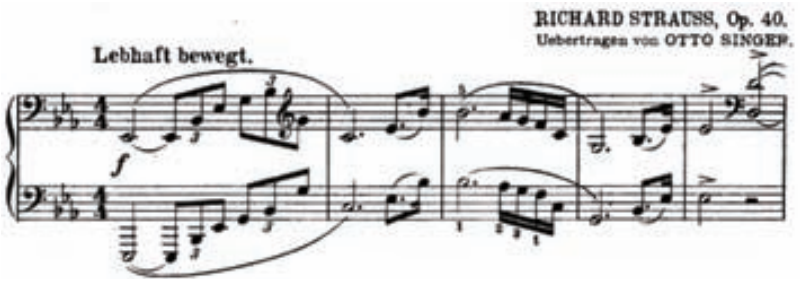
Regarding Methodology, the same as those mentioned in the previous chapter:

- Trajectory Overlapping Analysis –involving *omega* and specially *lambda* indexes- in combination with the Total Number of Conformations Analysis has proved to be a powerful MD conformational search tool for determining the Saturation when exploring the Conformational Space.
- It has been detected an MD dependence on starting conformations, meaning that MD trajectories tend to explore areas close to the starting position. It has been already suggested to be considered as a timescale problem.

Regarding Cyclodextrins:

- Flexibility grows as the number of glucoses grows.
- Descriptor D-II is a good choice for small cyclodextrins. However, it is not the appropriate one for large CDs because it is unable to model folding phenomena.
- It is necessary to develop a new descriptor –more general and less precise- that correctly describes folding and the secondary structure.
- The solvent influence can be measured qualitatively and quantitatively.

CONCLUSIONS



Richard Strauss
A Hero's Life op. 40
Tone Poem for Large Orchestra
(1899)
The Hero (Leitmotiv)

"Long was the path to arrive at this point..."

8 CONCLUSIONS

Having presented the results of the whole work, we are in position to come back to the first chapter, state again the objectives of this Thesis, and give the best answers to them:

8.1 Question I

The inherent problem regarding Molecular Dynamics inefficient exploration of the conformational space suggested us the necessity of developing a general methodology in the area of knowledge of "Conformational Search and MD analysis" as the first objective of this work.

8.1.1 Conclusion I

The Methodology proposed in the present work, which includes the development of the following tools: 1) A univocal 3D Molecular Descriptor; 2) The Saturation Analysis for Conformational Search; 3) The Trajectories Overlapping Ratio Analysis, has proved to be useful and adequate when applied to our research and, actually, it can be considered as an efficient good one.

8.2 Question II

As we use a full family of cyclodextrins (CDs) as a benchmark to evaluate the efficiency of our methodology, the choice of these molecules will allow us to give a proper answer to the "second objection" to Dr. Beà's work: "The data related to Molecular Dynamic (MD) calculations seemed to show a tight dependence on starting conformations and, apparently, it leads to the conclusion that the exploration of the conformational space of the systems under study were inefficiently done. This –meaning also a lack of generality- should be revised in order to assure the validity of the conclusions."

8.2.1 Conclusion II

As the referee stated in his replica, Molecular Dynamics does inefficiently explore the Conformational Space, especially when referring to large systems.

APPENDIX

Johann Sebastian Bach
The Musical Offering BWV 1079
(Collection of canons, fugues and other pieces of music
based on a musical theme by Frederick II of Prussia)
(1747)
Ricercar a 6 (six-voice fugue)

"The first step in the acquisition of wisdom is silence, the second listening, the third memory, the fourth practice, the fifth teaching others."

Solomon Ibn Gabirol

9 APPENDIX: METHODOLOGY

9.1 INTRODUCTION

As said in the Introduction, understanding Molecular Modelling is nowadays almost impossible not mentioning Classical Mechanics, Quantum Mechanics and Computer Science. However, just the most important areas directly related to the research developed in the present Thesis are being shown in this Chapter. For a wider and general insight, the consultation of the reference book in the area by Andrew R. Leach: *Molecular Modelling; Principles and Applications*, already cited in the Introduction[81], is highly recommendable.

9.2 MOLECULAR MODELLING

Many of the available computational methodologies[202] successfully model the behaviour of molecules and systems. Most of them –the *ab initio* techniques, the DFT[203] (Density Functional Theory) calculations, the Semi-Empirical and combined QM/MM approaches- have in common their relationship to the Schrödinger Wave Equation [Eq. 9.1 and 9.2]:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\Psi(\vec{r},t) = -\frac{\hbar}{i}\frac{\partial}{\partial t}\Psi(\vec{r},t)$$

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\psi(\vec{r}) = E\psi(\vec{r}) \Rightarrow \hat{H}\psi(\vec{r}) = E\psi(\vec{r})$$

Equations 9.1 and 9.2: Eq. 9.1 is the time-dependent equation while Eq. 9.2 is the time-independent version (also in the Eigenvalues formalism).

Nevertheless, we are focusing our interest *exclusively on conformational changes* of macromolecules. Therefore, any computational treatment involving molecular bonding, electronic effects or reactivity becomes absolutely unnecessary and means a waste of computational time. This level of calculation is not only unaffordable for our systems[204] but also inadequate.

[202] (a) Levine, I.N.; *Química Cuántica*. Editorial AC. 1ª Ed. En Español. **1977**. ISBN 84-7288-014-1. (b) Bertrán Rusca, J.; Branchadell Gallo, V.; Moreno Ferrer, M.; Sodupe Roure, M.; *Química Cuántica*. Editorial Síntesis S.A. **2000**. ISBN 84-7738-742-7.
[203] Geerlings, P.; De Proft, F.; Langenaeker, W.; *Chem. Rev*. **2003**. *103(5)*. 1793-1874.
[204] Lipkowitz, K.B.; *Chem. Rev.* **1998**. *98(5)*. 1829-1873.

9.3 MOLECULAR MECHANICS

A much more suitable methodology according to our necessities is employed here: that of empirical methods based on Molecular Mechanics[205], closely related to Newton's Second Law [Eq. 9.3]:

$$\sum_j \vec{f}_j = \vec{F}_i = m_i \frac{\partial^2 \vec{r}_i}{\partial t^2}$$

Equation 9.3

As said, Molecular Mechanics software packages are fully based on classical equations rather than quantum-mechanical ones searching for optimal computational performance.

9.3.1 Classical Equations

Different sets of equations are employed to model the vibrational and rotational molecular movements –the last ones being responsible for conformational changes- [Eq. 9.4].

$$E_{Total} = E_s + E_b + E_t + E_{it} + E_{vdw} + E_{elec} + E_{HBond} + E_{CrossTerms} + E_{solv} + \ldots$$

Equation 9.4

In this sense, Taylor series expansions –centred at the equilibrium points- are the best choice for approaching a potential energy profile [Eq. 9.5]:

$$V(q) = V(q)\big|_{q_0} + \sum_{i=1}^{3n} \left(\frac{\partial V(q)}{\partial q} \right)\Bigg|_{q_0} \Delta q_i + \frac{1}{2!} \sum_{i,j=1}^{3n} \left(\frac{\partial^2 V(q)}{\partial q_i \partial q_j} \right)\Bigg|_{q_0} \Delta q_i \Delta q_j + \frac{1}{3!} \sum_{i,j,k=1}^{3n} \left(\frac{\partial^3 V(q)}{\partial q_i \partial q_j \partial k_k} \right)\Bigg|_{q_0} \Delta q_i \Delta q_j \Delta q_k + \ldots$$

Equation 9.5

In which the zero-order term –the energy at the equilibrium point- is arbitrarily set to zero, and the first-order term vanishes because the derivative at the equilibrium point –the force- is zero. Therefore, the expansion, all in all, comprises second-order terms and others above [Eq. 9.6]:

[205] Burkert, U.; Allinger, N.L.; *Molecular Mechanics. ACS Monograph 177*. Ed. American Chemical Society. **1982**. ISBN 0-8412-0584-1.

$$V(q-q_0)=\frac{1}{2!}\left(\frac{d^2V(q)}{dq^2}\right)\Bigg|_{q=q_0}(q-q_0)^2+\sum_{n=3}O(q-q_0)^n$$

Equation 9.6

There are many possibilities, and different software packages use different series to approach the molecular behaviour. Particularly AMBER[206], the software mainly employed in our research, takes the simplest functional forms that preserve the essential nature of molecules in condensed phases –only second order terms when possible- to improve computational performance ratio [Eq. 9.7 to 9.13]:

Stretching: Hooke's Law. **Equation 9.7**

$$E_s=\frac{1}{2}K_r(r-r_0)^2$$

Bending: Hooke's Law. **Equation 9.8**

$$E_b=\frac{1}{2}K_\theta(\theta-\theta_0)^2$$

Torsion: Series of cosines. **Equation 9.9**

$$E_t=\frac{1}{2}\left[V_1(1+\cos\omega)+V_2(1-\cos 2\omega)+V_3(1+\cos 3\omega)\right]$$

Improper Torsion: Hooke's Law. **Equation 9.10**

$$E_{it}=\frac{1}{2}K_\xi(\xi-\xi_0)^2$$

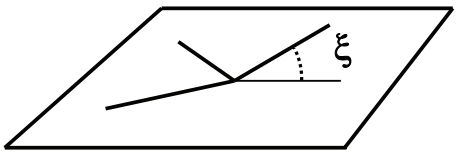


Figure 9.1: Example of improper torsion.

Electrostatic: Coulombic-potential. **Equation 9.11**

$$E_{Coulomb}=\frac{1}{4\pi\varepsilon_r}\frac{q_iq_j}{r_{ij}}$$

[206] (a) Pearlman, D.A.; Case, D.A.; Caldwell, J.W.; Ross, W.S.; Cheatham III, T.E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P.A.; *Comp. Phys. Commun*. **1995**. *91(1-3)*. 1-41. (b) Case, D.A.; Cheatham III, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, Jr., K.M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods. R.J.; *J. Comput. Chem*. **2005**. *26(16)*. 1668-1688. (c) Case, D.A.; Pearlman, D.A.; Caldwell, J.W.; Cheatham III, T.E.; Wang, J.; Ross, W.S.; Simmerling, C.L.; Darden, T.A.; Merz, K.M.; Stanton, R.V.; Cheng, A.L.; Vincent, J.J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.J.; Duan, Y.; Pitera J.; Massova, I.; Seibel, G.L.; Singh, U.C.; Weiner, P.K.; Kollman, P.A.; **2002**. **AMBER 7**. University of California, San Francisco.

Van der Waals: Lennard-Jones potential. **Equation 9.12**

$$E_{vdw} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

Hydrogen Bonding: Lennard-Jones potential. **Equation 9.13**

$$E_{HBond} = \frac{\alpha_{ij}}{r_{ij}^{12}} - \frac{\beta_{ij}}{r_{ij}^{6}}$$

Therefore, the general equation for a molecular system in the molecular mechanics approach is that of [Eq. 9.14]:

$$V = \sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + \cos\left(n\omega + \gamma\right)\right] + \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon_r r_{ij}} \right)$$

Equation 9.14

And, if atom-centred dipole polarizabilities are considered, an extra term including polarization is added to the equation above [Eq. 9.15]:

Polarization: $$E_{pol} = -\frac{1}{2}\sum_{i}^{atoms} \mu_i \cdot E_i^{(0)}$$ **Equation 9.15**

9.3.2 Force Fields

The full set of self-consistent parameters and constants necessary to model the individual atomic movements described by the equations shown above is called the *Force Field*. These values account for the energies, equilibrium distances, angles and dihedrals necessary to describe the molecular behaviour. The total number of parameters in the list depends on the number and type of atomic interactions (bending, stretching, torsion…) and also on the number of polynomial terms in the equations.

9.3.2.1 List of Force Fields

A great variety of Force Fields have been developed over the last years in order to model specific systems, gradually achieving different levels of success:

- MM2[207], MM3[208], and MM4[209] (Allinger): Organic molecules.
- MMFF[210] (Halgren): Organic molecules (oriented to drug design).
- OPLS[211] (Jorgensen): Liquids.
- ECEPP[212] (Scheraga): Peptides.
- AMBER[213]. (Kollman):
    - parm94[214] and parm99[215]: Proteins, Nucleic Acids and Biomolecules.
    - gaff[216]: Organic molecules and Biomolecules in general.
- TRIPOS 5.2[217] (Clark): Organic molecules and Biomolecules in general.
- CHARMM[218] (Karplus): Proteins and Lignin[219].
- GLYCAM[220] (Woods): Carbohydrates and Glycoproteins.
- GROMOS[221] (Hermans): Carbohydrates.
- HSEA[222] (Thogersen): Carbohydrates.

[207] Allinger, N.L.; *J. Am. Chem. Soc*. **1977**. *99(25)*. 8127-8134.
[208] (a) Allinger, N.L.; Yuh, Y.H.; Lii, J.-H.; *J. Am. Chem. Soc*. **1989**. *111(23)*. 8551-8566. (b) Allinger, N.L.; Lii, J.-H.; *J. Am. Chem. Soc*. **1989**. *111(23)*. 8566-8575. (c) Allinger, N.L.; Lii, J.-H.; *J. Am. Chem. Soc*. **1989**. *111(23)*. 8576-8582.
[209] (a) Allinger, N.L.; Chen, K.; Lii, J.-H.; *J. Comput. Chem*. **1996**. *17(5-6)*. 642-668. (b) Nevins, N.; Chen, K.; Allinger, N.L.; *J. Comput. Chem*. **1996**. *17(5-6)*. 669-694. (c) Nevins, N.; Lii, J.-H. Allinger, N.L.; *J. Comput. Chem*. **1996**. *17(5-6)*. 695-729. (d) Nevins, N.; Allinger, N.L.; *J. Comput. Chem*. **1996**. *17(5-6)*. 730-746. (e) Allinger, N.L.; Chen, K.; *J. Comput. Chem*. **1996**. *17(5-6)*. 747-755.
[210] (a) Halgren, T.A.; *J. Comput. Chem*. **1996**. *17(5-6)*. 490-519. (b) Halgren, T.A.; *J. Comput. Chem*. **1996**. *17(5-6)*. 520-552. (c) Halgren, T.A.; *J. Comput. Chem*. **1996**. *17(5-6)*. 553-586. (d) Halgren, T.A.; Nachbar, R.B.; *J. Comput. Chem*. **1996**. *17(5-6)*. 587-615. (e) Halgren, T.A.; *J. Comput. Chem*. **1996**. *17(5-6)*. 616-641. (f) Halgren, T.A.; *J. Comput. Chem*. **1999**. *20(7)*. 720-729. (g) Halgren, T.A.; *J. Comput. Chem*. **1999**. *20(7)*. 730-748.
[211] Jorgensen, W.L.; Tirado-Rives, J.; *J. Am. Chem. Soc*. **1988**. *110(6)*. 1657-1666.
[212] (a) Momany, F.A.; Carruthers, L.M.; McGuire, R.F.; Scheraga, H.A.; *J. Phys. Chem*. **1974**. *78(16)*. 1595-1620. (b) Momany, F.A.; Carruthers, L.M.; Scheraga, H.A.; *J. Phys. Chem*. **1974**. *78(16)*. 1621-1630.
[213] (a) Weiner, S.J.; Kollman, P.A.; Case, D.A.; Chandra Singh, U.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P.; *J. Am. Chem. Soc*. **1984**. *106(3)*. 765-784. (b) Kollman, P.A.; *Acc. Chem. Res*. **1996**. *29(10)*. 461-469. (b) Cheatham III, T.E.; Young, M.A.; *Biopolymers*. **2001**. *56(4)*. 232-256. (c) Ponder, J.W.; Case. D.A.; *Adv. Prot. Chem*. **2003**. *66*. 27-85.
[214] Cornell, W.D.; Cieplak, P.; Bayly, C.I.; Gould, I.R.; Merz, K.M.; Ferguson, D.M.; Spellmeyer, D.C.; Fox, T.; Caldwell, J.W.; Kollman, P.A.; *J. Am. Chem. Soc*. **1995**. *117(19)*. 5179-5197.
[215] Wang, J.; Cieplak, P.; Kollman, P.A.; *J. Comput. Chem*. **2000**. *21(12)*. 1049-1074.
[216] Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A.; *J. Comput. Chem*. **2004**. *25(9)*. 1157-1174.
[217] Clark, M.; Cramer III, R.D.; van Opdenbosch, N. *J. Comput. Chem*. **1989**. *10(8)*. 982-1012.
[218] Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S.; Karplus, M.; *J. Comput. Chem*. **1983**. *4(2)*. 187-217.
[219] Petridis, L.; Smith, J.C.; *J. Comput. Chem*. **2009**. *30(3)*. 457-467.
[220] Woods, R.J.; Dwek, R.A.; Edge, C.J.; Fraser-Reid, B.; *J. Phys. Chem*. **1995**. *99(11)*. 3832-3846.
[221] Hermans, J.; Berendsen, H.J.C.; van Gunsteren, W.F.; Postma, J.P.M.; *Biopolymers*. **1984**. *23(8)*. 1513-1518.
[222] (a) Lemieux, R.U.; Bock, K.; Delbaere, L.T.J.; Koto, S.; Rao, V.S. *Can. J. Chem*. **1980**. *58(6)*. 631-653. (b) Thogersen, H.; Lemieux, R.U.; Bock, K.; Meyer, B.; *Can. J. Chem*. **1982**. *60(1)*. 44-57.

- AMB99C[223] (Momany): Carbohydrates.
- CSFF[224] (Brady): Carbohydrates.

Further improvements and corrections were made to several Force Fields in the 90's when atomic-centred induced dipoles and polarisation phenomena in general were considered to be of importance[225]. They were called the *"non-additive"* Force Fields and included polarization terms or enhanced Potential Energy Surface (PES) parametrisation.

9.3.2.2 Parametrisation

Creating a Force Field involves basically two steps. On the one hand, a huge amount of experimental information –obtained by means of spectroscopic techniques (X-ray diffraction, IR-Raman, NMR…) or *ab initio* calculations- is gathered together. On the other hand, the raw data is processed using multivariate fitting techniques until the full set of self-consistent parameters can be derived.

The Force Field is then tested with molecules other –and larger- than those employed for the parametrisation to check its transferability[226]. The results are then compared against experimental data. The refinement process is iteratively done –by slightly adjusting already studied parameters and adding new ones not yet considered- and leads to an enhanced set of parameters.
Anyway, the inclusion of all these new terms responsible for the improvements in accuracy increases the length of the functional forms and directly affects the computational performance, so a compromise situation between both realities is required.

[223] (a) Momany, F.A.; Willett, J.L.; *Carbohydr. Res*. **2000**. *326(3)*. 194-209. (b) Momany, F.A.; Willett, J.L.; *Carbohydr. Res*. **2000**. *326(3)*. 210-226.
[224] Kuttel, M.; Brady, J.W.; Naidoo, K.J.; *J. Comput. Chem*. **2002**. *23(13)*. 1236-1243.
[225] (a) Ferenczy, G.G.; *J. Comp. Chem*. **1991**. *12(8)*. 913-917. (b) Chipot, C.; Angyan, J.G.; Ferenczy, G.G.; Scheraga, H.A.; *J. Phys. Chem*. **1993**. *97(25)*. 6628-6636. (c) Winn, P.J.; Ferenczy, G.G.; Reynolds, C.A.; *J. Phys. Chem. A*. **1997**. *101(30)*. 5437-5445. (d) Ferenczy, G.G.; Winn, P.J.; Reynolds, C.A.; *J. Phys. Chem. A*. **1997**. *101(30)*. 5446-5455. (e) Winn, P.J.; Ferenczy, G.G.; Reynolds, C.A.; *J. Comput. Chem*. **1999**. *20(7)*. 704-712.
[226] Ivanov, P.; *Lecture Notes in Molecular Mechanics*. Ed. Universitat Autònoma de Barcelona. Barcelona. **1996**.

The Force Field parametrisation is not a second rate topic in biomolecular chemistry when investigating structure-activity relationships[227]. Just considering the list given above, we realise that the number of those which are *"specifically addressed"* to carbohydrates points that all of them have weak and strong points, and none is *"perfect"*. Therefore, evaluating force fields[228] and trying to guess which one is the most suitable for certain macromolecular system is also a field of study and the frame in which Dr. Itziar Maestre centred her research, as said in the Objectives.

The principal Force Field employed in the present Thesis when AMBER 7 was used has been *parm99* –the originally one selected by Dr. Beà in his research-, although MM3* was selected when GACK-MacroModel calculations were done. Merz-Kollman set of charges[229] were employed for the units in the AMBER 7 calculations.

9.3.3 Solvation Models

The importance of working with adequate Force Fields has been just explained. As seen, they mainly describe equations that account for molecular movements –like vibrations and electrostatics- within the macromolecule itself, although this is an incomplete point of view.

We know that most of chemical reactions take place in solution and the effect of solvent is determinant[230]. It is the case of large macromolecules –proteins, fibres, polymers…- whose folding properties and chemical and biological activities strongly depend on the conformation they adopt. Therefore, solvent models that accurately take into account substrate-solvent interactions prove to be of importance.

[227] MacKerell Jr., A.D.; *J. Comput. Chem*. **2004**. *25(13)*. 1584-1604.
[228] Hemmingsen, L.; Madsen, D.E.; Esbensen, A.L.; Olsen, L.; Engelsenc, S.B.; *Carbohydr. Res.* **2004**. *339(5)*. 937-948.
[229] Singh, U.C.; Kollman, P.A.; *J. Comput. Chem*. **1984**. *5(2)*. 129-145.
[230] (a) Orozco, M.; Luque, F.J.; *Chem. Rev*. **2000**. *100(11)*. 4187-4226. (b) Tomasi, J.; Mennucci, B.; Cammi, R.; *Chem. Rev*. **2005**. *105(8)*. 2999-3094.

9.3.3.1 Implicit Solvation Models

The principal handicap found by software developers when they faced this problem was the huge amount of molecules necessary to create a bulk of solvent that behaves properly both in the microscopic and the macroscopic scale.

Three decades ago, computationally speaking, such a big system was absolutely unaffordable so researchers approached the problem in a simplified way. They only paid attention to the *electrostatic* part, avoiding the explicit inclusion of a *large number of particles* within the system, assuming that solvent effects were mainly originated by electrostatic and non-bonding interactions. The result was the so called *Continuum Solvation Models*[231], like the *Generalised Born*[232] (GB) [Eq. 9.16] and the P*oisson-Boltzmann*[233] (PB) [Eq. 9.17], where the first equation –Born- describes the PES as the result of a summation of all coulombic interactions between spherical atoms, and the second one –Poisson-Boltzmann- does it generating a continuous curvilinear surface.

$$G_{GB} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon_r}\right)\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j \exp\left(-r_{ij}^2/4a_i a_j\right)}}$$

$$\nabla\left[\varepsilon(r)\nabla\phi(r)\right] - \kappa\sinh\left[\phi(r)\right] = -4\pi\rho(r)$$

Equations 9.16 and 9.17

In both cases, the solvent effect was easily implemented in the calculation just by introducing the macromolecule into a continuum environment whose dielectric constant, ε_r, equals that of the solvent [Fig. 9.2]. The computational requirements were low and this solution was quite successful.

[231] (a) Tomasi, J.; Persico, M.; *Chem. Rev*. **1994**. *94(7)*. 2027-2094. (b) Cramer, C.J.; Truhlar, D.G.; *Chem. Rev*. **1999**. *99(8)*. 2161-2200.
[232] (a) Bashford, D.; Case, D.A.; *Annu. Rev. Phys. Chem*. **2000**. *51*. 129-152. (b) Feig, M.; Im, W.; Brooks III, C.L.; *J. Chem. Phys*. **2004**. *120(2)*. 903-911. (c) Schaefer, M.; Karplus, M.; *J. Phys. Chem. B*. **1996**. *100(5)*. 1578-1599.
[233] (a) Davis, M.E.; McCammon, J.A.; *Chem. Rev*. **1990**. *90(3)*. 509-521. (b) Honig, B.; Nicholls, A.; *Science*. **1995**. *268*. 1144-1149.

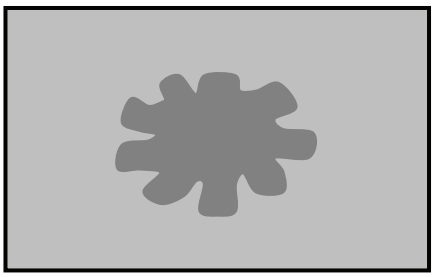


Figure 9.2: Molecule immerse in a continuum solvent model.

The next improvement came when modellers realised that the effect of the solvent cannot be equally treated all over the molecular surface. This correction was added because parts of certain molecules –especially macromolecules- remain partially hidden and prevent solvent molecules to freely access them because of the steric hindrance. Those parts, sometimes called *pockets*, are not in contact with solvent and therefore they do not feel the electrostatic effects. The *Solvent Accessible Surface*[234] (SAS) was introduced confronting to *Molecular Surface*. SAS is evaluated by increasing the radius of a hypothetical water –solvent- sphere rolling over the van der Waals molecular surface. The actual surface spanned by the centre of such a rolling sphere describes the SAS [Fig. 9.3].

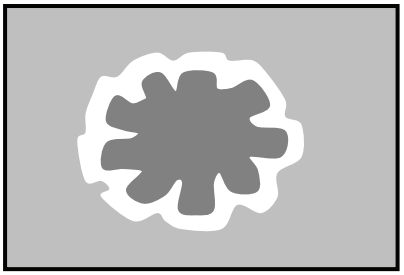


Figure 9.3: Molecule immerse in a continuum solvent model. The area delimited in white represents the portion of space not accessible to the solvent.

New corrections were also included[235]: one that took into account the energy associated to the solvent-solute non-polar interactions, G_{vdW}, and another one that estimated the solvent rearrangement entropy to room space for the solute, G_{cav}. The results were the GB/SA[236] [Eq. 9.18] and the PB/SA[237] solvent models.

[234] (a) Lee, B.; Richards, F.M.; *J. Mol. Biol*. **1971**. *55(3)*. 379-400. (b) Scarsi, M.; Apostolakis, J.; Caflisch, A.; *J. Phys. Chem. A*. **1997**. *101(43)*. 8098-8106.
[235] Still, W.C.; Tempczyk, A.; Hawley, R.C.; Hendrickson, T.; *J. Am. Chem. Soc*. **1990**. *112(16)*. 6127-6129.
[236] Mitomo, D.; Watanabe, Y.S.; Kamiya, N.; Higo, J.; *Chem. Phys. Lett*. **2006**. *427(4-6)*. 399-403.
[237] Fogolari, F.; Brigo, A.; Molinari, H.; *J. Mol. Recognit*. **2002**. *15(6)*. 377-392.

$$G_{SA} = G_{vdW} + G_{cav} = \sum_k \sigma_k SA_k$$

Equation 9.18: Solvation free energy (G_{SA}) consists of the summation of a solvent-solvent cavity term (G_{cav}), and a solute-solvent *van der Waals* term (G_{vdW}), where SA_k is the total solvent accessible surface area of atoms of type "k" and "σ_t" is an empirical atomic solvation parameter.

Despite the refinements, continuum models were still inappropriate to effectively describe the solvent-solute interactions because of their extreme simplicity[238]. In fact, solvent effects go beyond electrostatic and non-polar interactions, including also body collisions and complex dynamical behaviour such as *Brownian motion*[239], being the latest almost impossible to be reproduced by means of continuum environments. Stochastic Dynamics (SD) was developed to include that effect by means of the Langevin Equation[240] [Eq. 9.19] –although it usually involves previous explicit solvent MD calculations to obtain the collision frequency, γ_i, and then the friction coefficient, ξ_i-.

$$m_i \frac{dv_i}{dt} = F\{x_i(t)\} - \gamma_i m_i v_i + R_i(t)$$

Equation 9.19: Langevin Equation: The first term, F{x(t)}, accounts for the interaction between the particle and other particles. The second term, γmv, arises from the motion of the particle through the solvent and represents the frictional drag on the particle due to the solvent. The third term, R(t), represents the random fluctuations caused by interactions with solvent molecules.

9.3.3.2 Explicit Solvation Models

The new solvent models were more ambitious and consisted of modelling a *real solvent box* that included a large amount of *rigid* solvent molecules surrounding the solute [Fig. 9.4]. This proposal was still, computationally speaking, exceptionally time-consuming; however processors were at that time powerful enough to allow such calculations, especially when they ran under periodic boundary conditions at parallel supercomputers.

[238] Zhou, R.; Berne, BJ.; *Proc. Nat. Acad. Sci. USA*. **2002**. *99(20)*. 12777-12728.
[239] (a) Brown, R.; *Phil. Mag*. **1828**. *4*. 161-173. (b) Einstein, A.; *Ann. Phys*. **1905**. *17*. 549-560. Reprint [1956] from the English translation [1926] from the original paper in German. (c) von Smoluchowsky, M.; *Ann. Phys*. **1906**. *21*. 756-780. [In German]. (d) Uhlenbeck, G.E.; Ornstein, L.S.; *Phys. Rev*. **1930**. *36(5)*. 823-841.
[240] Nadler, W.; Brunger, A.T.; Schulten, K.; Karplus, M.; P*roc. Nat. Acad. Sci. USA*. **1987**. *84(22)*. 7933-7937.

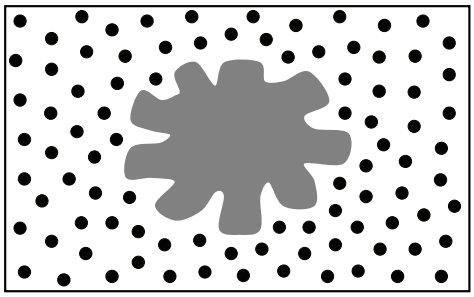


Figure 9.4: Molecule immerse in an explicit solvent model. The black dots represent the solvent molecules surrounding the macromolecule.

Since then, several models for the most common solvents were created and updated: Water is probably the one most extensively used. There are several models depending on the flexibility and the number of charges for the Coulombic interaction, i.e. TIP3P[241], TIP4P[241] and TIP5P[242], representing rigid monomers with 3, 4, and 5 interacting sites. The van der Waals force between intermolecular pairs of charges follows the Lennard-Jones equation centred in the Oxygen atom. Further models included also polarisation, TIP4P-Pol[243], and long distance enhanced Coulombic interactions, TIP4P-Ew[244].

Moreover, there are also a great variety of organic solvents and their parameters available, like ethanol[245], propanol[245], methanol[245,246], acetamide[247], N-methylacetamide[246], acetonitrile[248], acetic acid[249], methyl acetate[249], acetone[249], pyrimidine[249], chloroform[249,250], benzene[251], methylamine[252], methanethiol[253], toluene[254], phenol[254], and pyridine[254,255] and more.

In the present Thesis two explicit solvent models have been employed: TIP3P version of Water and *in home* BNZ version of Benzene.

[241] Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L.; *J. Chem. Phys*. **1983**. *79(2)*. 926-935.
[242] Mahoney, M.W.; Jorgensen, W.L.; *J. Chem. Phys*. **2000**. *112(20)*. 8910-8922.
[243] Chen, B.; Xing, J.; Siepmann, J.I.; *J. Phys. Chem. B*. **2000**. *104 (10)*. 2391-2401.
[244] (a) Horn, H.W.; Swope, W.C.; Pitera, J.W.; Madura, J.D.; Dick. T.J.; Hura, G.L.; Head-Gordon, T.; *J. Chem. Phys*. **2004**. *120(20)*. 9665-9678. (b) Mobley, D.L.; Dumont, E.; Chodera, J.D.; Dill, K.A.; *J. Phys. Chem. B*. **2007**. *111(9)*. 2242-2254.
[245] Jorgensen, W.L.; *J. Phys. Chem*. **1986**. *90(7)*. 1276-1284.
[246] Caldwell, J.W.; Kollman, P.A.; *J. Phys. Chem*. **1995**. *99(16)*. 6208-6219.
[247] Jorgensen, W.L.; Swenson, C.J.; *J. Am. Chem. Soc*. **1985**. *107(3)*. 569-578.
[248] (a) Grabuleda, X.; Jaime, C.; Kollman, P.A.; *J. Comput. Chem*. **2000**. *21(10)*. 901-908. (b) Nikitin, A.M.; Lyubartsev, A.P.; *J. Comput. Chem*. **2007**. *28(12)*. 2020-2026.
[249] Jorgensen, W.L.; Briggs, J.M.; Contreras, M.L.; *J. Phys. Chem*. **1990**. *94(4)*. 1683-1686.
[250] Jorgensen, W.L.; Boudon, S.; Nguyen, T.B.; *J. Am. Chem. Soc*. **1989**. *111(2)*. 755-757.
[251] Cacelli, I.; Cinacchi, G.; Prampolini, G.; Tani, A.; *J. Am. Chem. Soc*. **2004**. *126(43)*. 14278-14286.
[252] Jorgensen, W.L.; Briggs, J. M.; *J. Am. Chem. Soc*. **1989**. *111(12)*. 4190-4197.
[253] Jorgensen, W.L.; *J. Phys. Chem*. **1986**. *90(23)*. 6379-6388.
[254] Baker, C.M.; Grant, G.H.; *J. Chem. Theory Comput*. **2007**. *3(2)*. 530-548.
[255] Jorgensen, W.L.; *J. Am. Chem. Soc*. **1989**. *111(10)*. 3770-3771.

9.3.3.3 Periodic Boundary Conditions (PBC)

Periodic Boundary Conditions (PBC) are particularly useful for simulating bulk solvent systems. The methodology is iterative; once the solvent box containing one macromolecule has been set up, the cell is replicated in all three dimensions until the whole space is covered [Fig. 9.5].

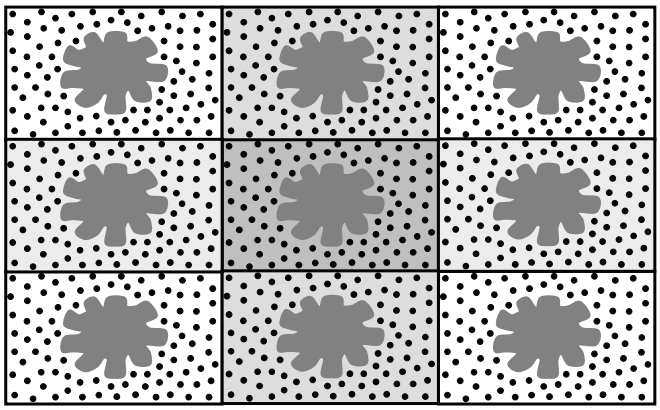


Figure 9.5: The cell in the middle is repeated in all directions; All the cells in the schema are exactly the same.

This behaviour is algorithmically achieved by allowing molecules and interactions to access the cell by the opposite face when they exit the box from a given side [Fig. 9.6].

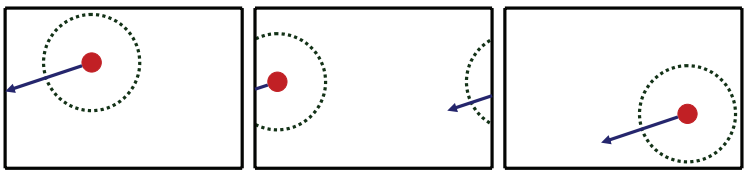


Figure 9.6: The continuity is achieved by allowing molecules and interactions pass from one side to the other side.

However, PBC also introduces correlational artefacts that do not respect the translational invariance of the system –and therefore affects the trustworthy of the macroscopic properties- when parameters are unwisely selected[256], especially when involving *cutoff* values[257] [Fig. 9.7].

[256] Shirts, R.B.; Burt, S.R.; Johnson, A.M.; *J. Chem. Phys*. **2006**. *125(16)*. 164102.
[257] (a) Schreiber, H.; Steinhauser, O.; *Biochemistry*. **1992**. *31(25)*. 5856-5860. (b) de Souza, O.N.; Ornstein, R.L.; *Biophys. J*. **1997**. *72(6)*. 2395-2397. (c) Cheatham III, T.E.; Miller, J.L.; Fox, T.; Darden, T.A.; Kollman, P.A.; *J. Am. Chem. Soc*. **1995**. *117(14)*. 4193-4194.

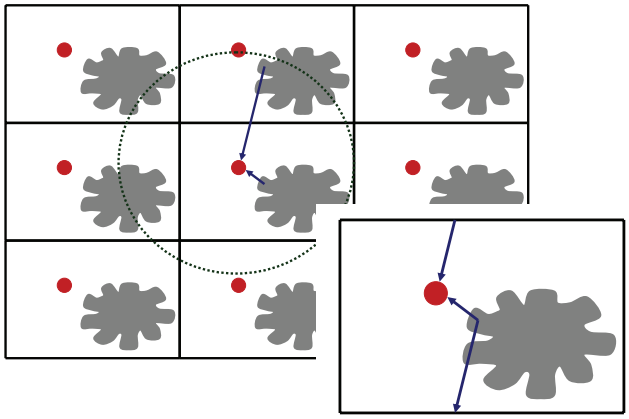
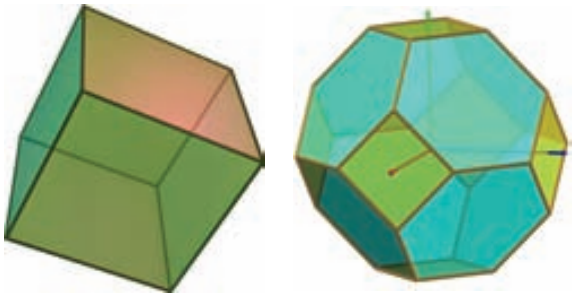


Figure 9.7: The figure depicts an "artefact". When *cutoff* values comparable to the dimensions of the cell are selected, the same solvent molecule interacts with the same macromolecular fragment – and *vice versa*- twice; from one side of the box and from the other both at the same time.

There are several possibilities of space-filling geometrical bodies and therefore can act as periodic cells. Anyway the most commonly used are the cube and the truncated octahedron [Fig. 9.8].



(a) **(b)**

Figure 9.8: (a) Cubic[258] and (b) Truncated Octahedral[259] polyhedron. The second one helps avoiding the *corner effects* and therefore improves solvent bulk behaviour.

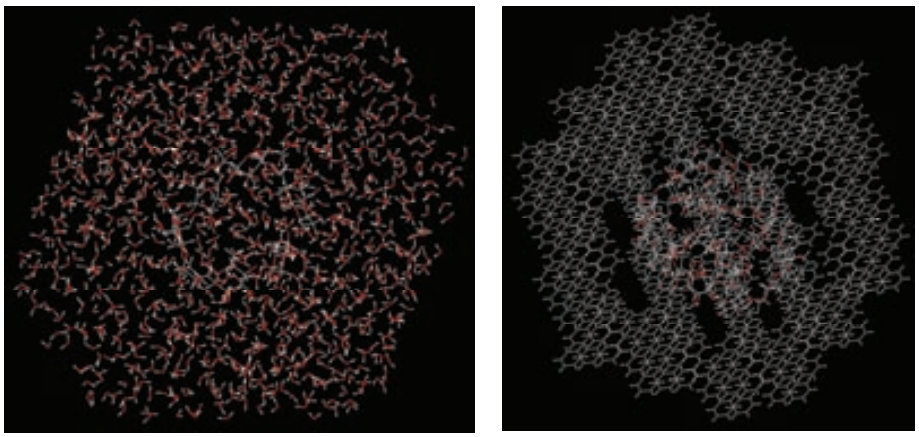
In the present Thesis, truncated octahedral solvent boxes under PBC have been employed whenever solvent simulations were done.

9.3.3.4 Mixed Solvation Models: MM-PBSA

MM-PBSA is a mixed solvation model proposed by Kollman and coworkers[260] included into the AMBER software package. It combines an explicit molecular mechanical model for the solute with a continuum method for the solvation free energy[261].

[258] Image source: http://es.wikipedia.org/wiki/Archivo:Hexahedron.jpg
[259] Image source: http://www.mathcurve.com/polyedres/octaedre_tronque/octaedre_tronque.shtml

The first step in a MM-PBSA calculation is obtaining a MD trajectory of the macromolecule inside an explicit solvent box [Fig. 9.9]. Then the solvent is fully removed and the trajectory alone computed according to the PBSA continuum model to evaluate the solvation free energy.



(a) **(b)**

Figure 9.9: Example of truncated octahedral explicit solvent boxes used in this Thesis: (a) gamma-CD in TIP3P water model and (b) CD28 in benzene.

In the present Thesis, MM-PBSA methodology has been used to evaluate the solvation energy of cyclodextrin-solvent systems.

9.3.3.5 Parameterisation

Developing a new explicit solvent model is in certain way similar to creating a Force Field.

The first step is defining all the parameters of the solvent molecule –distances, angles, dihedrals- and also calculating the charges on atoms of any interacting site –responsible for the electrostatic interactions- and the Lennard-Jones parameters –that account for the van der Waals interactions-. The second step is creating a solvent box with it and running long series of Molecular Dynamics calculations to evaluate the bulk

[260] Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, T.E.; *Acc. Chem. Res*. **2000**. *33(12)*. 889-897.
[261] Fogolari, F.; Brigo, A.; Molinari, H.; *Biophys. J.* **2003**. *85(1)*. 159-166.

macroscopic properties –like density, heat of vaporization, and isothermal compressibility- that emanate from those *microscopic parameters*.

Then, as in the case of Force Fields, theoretical results are compared against experimental data and the refinement process is iteratively done by adjusting charges and van der Waals parameters until full convergence between experimental and theoretical data is achieved.

Dr. Grabuleda, former member of our group, developed a solvent model for acetonitrile[248] during his PhD; likewise a benzene solvent model was also semi-developed in the present Thesis.

9.4 METHODOLOGIES

This section exposes a brief introduction to the most relevant methodologies used in this Thesis: Geometrical Optimization, Genetic Algorithms, Simulated Annealing / Monte Carlo, and Molecular Dynamics.

9.4.1 Geometrical Optimization

Geometrical Optimization –also known as Energy Minimization or simply Minimization- is not a Combinatorial Optimization Methodology and has nothing to do with Conformational Search in the wide sense we are dealing with here. However, this is a helpful side-tool to ensure full convergence to a minimum in combination with those Conformational Search techniques.

As said in the Introduction, it is a methodology based on Differential Calculus developed to obtain the closest minimum to the starting conformation [Fig. 9.10]. In this sense, this methodology is absolutely dependent on the starting structure and therefore cannot be trusted by itself as a Conformational Search Tool.

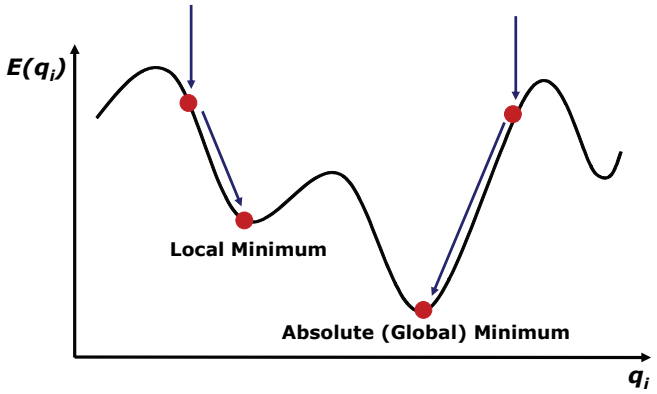


Figure 9.10: Schematic potential hypersurface with local and absolute minima. Minimization drives the system towards the minimum closest to the starting conformation.

There are several algorithms in the classical bibliography[262]: i.e. non-derivative methods (*simplex method*[263], *sequential univariate method*), first derivative methods (*steepest descent*[264]: *line search in one dimension*, or *arbitrary step approach*; and *conjugate gradients minimization*[265]), and second derivative methods (*Newton-Raphson*[266], *Block-diagonal Newton-Raphson*, and *Quasi-Newton methods*).

All of them gradually change the input coordinates to produce configurations with lower energies until the minimum is located. The termination criterion can be *structural* or *energetic*, and also *derivative* or *non-derivative*. Anyway, all the programs permit also to specify a maximum number of steps before stopping the iterations.
In the present thesis, both *conjugated gradient* and *steepest descent* methods have been employed in molecular mechanics calculations: AMBER 7 default protocol specifies ten steps of steepest descent at the beginning and then changes to conjugated gradient while the Macromodel default option is the Polak-Ribiere conjugated gradient for the whole process. Both of them use derivative termination criteria. However, quantum mechanics calculations done with Gaussian98[267] employed *second derivative algorithms* and four

[262] Bernhardt Schlegel, H.; *Encyclopedia of Computational Chemistry*. Vol. 2. 1136-1157. Ed. John Wiley & Sons. **1998**. ISBN 0-471-96588-X.
[263] Dantzig, G.; Orden, A.; Wolfe, P.; *Pacific Journal of Mathematics*. **1955**. *5(2)*. 183-195.
[264] Snyman J.A.; *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*. Springer Publishing. **2005**. ISBN 0-387-24348-8
[265] (a) Hestenes, M.R.; Stiefel, E.; *Journal of Research of the National Bureau of Standards*. **1952**. *49(6)*. 409-436. (b) Polak, E.; Ribière, G.; *Rev. Francaise Informat. Recherche Operationelle*. **1969**. *3e Année 16*. 35-43. (c) Fletcher, R.; Reeves, C.M.; *Computer Journal*. **1964**. *7(2)*. 149-154.
[266] Ypma, T.J.; *SIAM Review*. **1995**. *37(4)*. 531-551.
[267] **Gaussian 98**, *Revision A.11*.; Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Zakrzewski, V.G.; Montgomery, Jr., J.A.; Stratmann, R.E.; Burant, J.C.; Dapprich, S.; Millam, J.M.; Daniels, A.D.; Kudin, K.N.; Strain, M.C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G.A.; Ayala,

termination criteria[268]: forces = *cutoff*; RMS forces = *cutoff*; coordinates displacement = *cutoff*; RMS coordinates displacement = *cutoff* (by default value for the *cutoffs*).

9.4.2 Global Search Methods and Pool of Conformations

In order to locate more than one minimum –or the global energy minimum- methods for generating different starting points are required, and every one of them is minimised later to get a stable minimum. As said, specialised minimisation methods can go downhill and reach the closest energetic minimum to the input conformation but no algorithm in that group has yet proved capable of locating the global minimum from an arbitrary starting point.

This is a problem of *Combinatorial Optimization*, typical in chemistry, and other methods have to be used to approach it[269]. The next sections give a brief explanation of those that were mainly used in the present Thesis; *Genetic Algorithms* (GA), *Simulated Annealing* (SA), and *Molecular Dynamics* (MD).

9.4.2.1 The Scale-factor in conformational problems

Molecular size affects computing performance in several ways: It is known that there exist a relationship between the number of atoms in the molecule and the simulation time, which is proportional to the dimension of the matrix of interactions of the system.

Nevertheless, there are two more important properties related to molecular size whose implications seriously affect conformational search techniques: *topology* and *folding*

P.Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J.J.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Cioslowski, J.; Ortiz, J.V.; Baboul, A.G.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R.L.; Fox, D.J.; Keith, T.; Al-Laham, M.A.; Peng, C.Y.; Nanayakkara, A.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Andres, J.L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E.S.; and Pople, J.A.; **Gaussian, Inc**., Pittsburgh PA, **2001**.

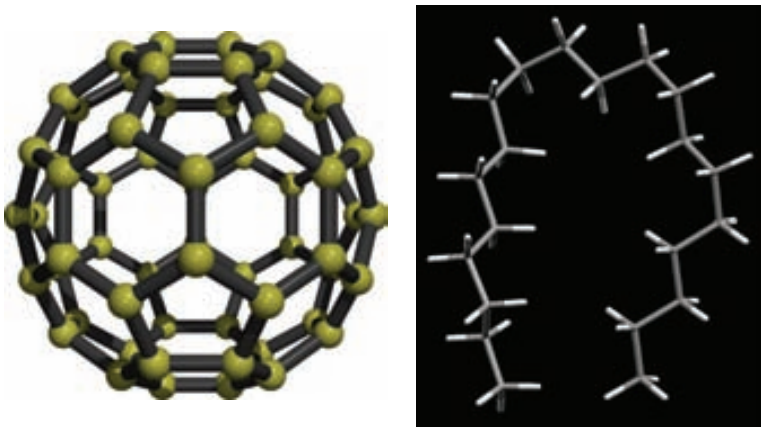
[268] Foresman, J.B.; Frisch, Æ.; *Exploring Chemistry with Electronic Structure Methods. 2nd Ed*. Gaussian, Inc. Pittsburgh, PA. **1996**. ISBN 0-9636769-3-8.

[269] (a) Saunders, M.; *J. Am. Chem. Soc*. **1987**. *109(10)*. 3150-3152. (b) Goto, H.; Osawa, E.; *J. Am. Chem. Soc*. **1989**. *111(24)*. 8950-8951. (c) Lipton, M.; Still, W.C.; *J. Comput. Chem*. **1988**. *9(4)*. 343-355. (d) Chang, C-.E.; Gilson, M.K.; *J. Comput. Chem*. **2003**. *24(16)*. 1987-1998. (e) Goodman, J.M.; Still, W.C.; *J. Comput. Chem*. **1991**. *12(9)*. 1110-1117. (f) Wang, C-.S.; *J. Comput. Chem*. **1997**. *18(2)*. 277-289. (g) Kolossvary, I.; Guida, W.C.; *J. Am. Chem. Soc*. **1993**. *115(6)*. 2107-2119. (h) Goto, H.; Osawa, E.; *J. Chem. Perkin. Trans. 2*. **1993**. *(2)*. 187-198. (i) Goodman, J.M.; Saz, A.B.; *J. Chem. Soc. Perkin. Trans. 2*. **1997**. *(6)*. 1201-1204. (j) Goodman, J.M.; Saz, A.B.; *J. Chem. Soc. Perkin. Trans. 2*. **1997**. *(6)*. 1205-1208.

properties. They not only drive the macromolecular behaviour but also force the researcher to use certain methodologies rather than others when investigating them.

9.4.2.1.1 Topology

Sometimes, the number of atoms in a molecule by itself has nothing to do with the expected conformational behaviour. I.e. the ordinary C_{60} fullerene gathers almost as many atoms as the lineal n-alkane $C_{20}H_{42}$ [Fig. 9.11], however, the conformational space of the alkane is immense –in the order of 3^{17}- in comparison to that of the fullerene –only 1 conformation-. In fact, any n-alkane having only a few carbons behaves similarly.



(a) **(b)**

Figure 9.11: (a) Fullerene and (b) Lineal n-alkane $C_{20}H_{42}$. Despite both of them have almost the same number of atoms, fullerene is rigid and n-alkane is extremely flexible.

So, topology is important. But probably the most important property, especially when dealing with proteins, polymers and also cyclodextrins, is folding.

9.4.2.1.2 Folding and Aggregation

The conformational space of small molecules is usually effectively explored by means of Monte Carlo, Low Mode Search and Genetic Algorithms methods. Most of them focus their perturbations on the torsional parameters because in small and medium-sized molecules, this sort of alterations enable full conformational changes when the variables

are wisely chosen. The key-point is that *local perturbations* spread all over the molecule rendering *total conformational changes*.

The situation is absolutely different when large macromolecules are involved, and it is not just a matter of number of atoms. As said in the Introduction, proteins, polymers, and cyclodextrins are large molecules that go beyond *primary structure*. They exhibit *secondary structure –folding-* [Fig. 9.12], responsible for some of their properties, and in the case of proteins, *tertiary* and even *quaternary structure* –aggregations-.

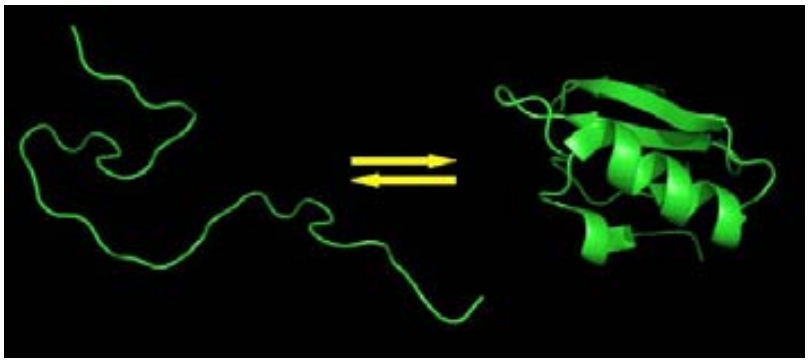


Figure 9.12: Image of unfolded / folded protein[270]. Secondary Structure is extremely important in proteins and polymers as it is responsible for their properties.

Local perturbations like that on torsional angles are unable to promote full conformational changes. In large macromolecules, they do not directly arise from dihedral alterations. All the electrostatic and non polar interactions responsible for the secondary structure stabilise the macromolecule and therefore it is necessary to previously unfold the molecule, apply the perturbations, and again let the system fold and relax into a new minimum.

We faced this problem in the present Thesis when the conformational space of cyclodextrins was studied. Monte Carlo was almost rejected at the very beginning, but it took time to realise that Genetic Algorithms implementation of perturbations would be also inappropriate. In the end, the most effective methodology proved to be Simulated Annealing in the thermal shock implementation precisely because it promoted unfolding and, therefore, full conformational changes.

[270] Image source: http://en.wikipedia.org/wiki/File:Protein_folding.png

9.4.2.2 Genetic Algorithms

Problems with non-deterministic solutions that run in polynomial time, like Conformational Search, are called NP-class problems[271]. Because of their high complexity they cannot be solved in a realistic timeframe using deterministic techniques. So, to solve these problems in a reasonable amount of time, heuristic methods must be used and Genetic Algorithms[272] (GA) are powerful ones, capable of efficiently searching large spaces of possible solutions.

This approach was originally proposed by Holland –who applied the evolutionary ideas to computational science- and popularised by Goldberg[273]. Its theoretical basis relies on the *Schema Theorem*.

9.4.2.2.1 Holland's GA Schema Theorem

This theorem is commonly explained in the binary version of the Canonical Genetic Algorithm (CGA). The properties of a CGA include:

- Binary alphabet.
- Fixed length, "l". Individuals –chromosomes- of equal length.
- Fitness proportional selection.
- Single point crossover.
- Gene wise mutation.

1) Binary Alphabet and Chromosomal Length

The first step in a CGA is formalising the variables –the set of *"genes"*- in your problem in the specific binary code to obtain your *"chromosome"*. Now, we need a few definitions to be explained:

[271] Schrijver, A.; *A Course in Combinatorial Optimization*. Ed. Department of Mathematics. University of Amsterdam. The Netherlands. **2008**.
[272] Haupt, R.L.; Haupt, S.E.; *Practical Genetic Algorithms*. 2nd Ed. John Wiley & Sons. **2004**. ISBN 0-471-45565-2.
[273] Goldberg, D.E.; *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley. **1989**. ISBN 0-201-15767-5.

- **Schema, *H***

"A schema is a subset of the space of all possible individuals –chromosomes- for which all the genes match the template for schema H".

If A denotes the alphabet of gene alleles, then $A \cup *$ is the schema alphabet, where * is the 'wild card' symbol matching any allele value.

I.e. For the binary alphabet: $A \in \{0,1,*\}$ where $* \in \{0,1\}$.

I.e. For a binary individual with the gene sequence {0 1 1 1 0 0 0}, then it follows that one –of many- matching schema might have the form, *H* = [* 1 1 * 0 * *].

I.e. The schema *H* = [0 1 * 1 *] identifies the chromosome set:

0 1 0 1 0

0 1 0 1 1

0 1 1 1 0

0 1 1 1 1

- **Schema Order, *o(H)***

"S*chema order, o(H), is the number of non '*' genes in schema H".*

I.e. *o*(* 0 * 0 1 * *) = 3.

- **Schema Defining Length, *δ(H)***

"Schema Defining Length, δ(H), is the distance between first and last non '' gene in schema H".*

I.e. *δ*(* 1 * 0 1 * *) = 5 – 2 = 3.

2) Fitness Proportional Selection

GA involves two opposed driving forces –represented by *operators*- that control the full process. On the one hand, there is the ***Selection Operator***; the "deterministic" force that makes the system to evolve towards the best solutions. This tends to select the best individuals in every generation and usually reduces the diversity of the conformational pool. However, it does nothing to improve the fitness of individuals in a population and also tends to get the system stuck in a local minimum [Fig. 9.13].

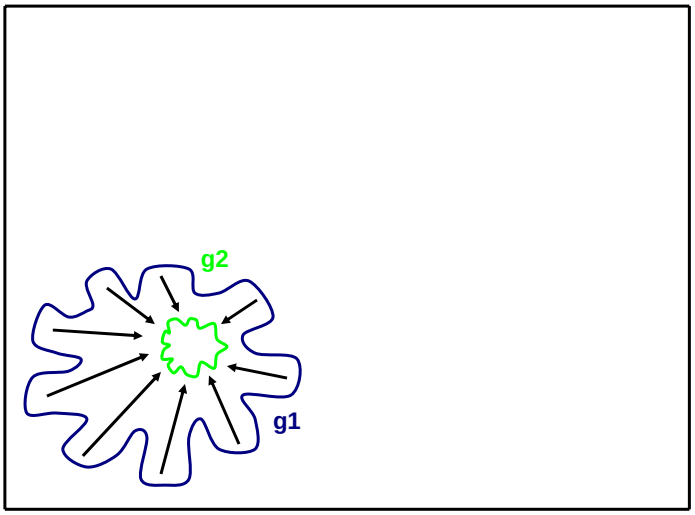


Figure 9.13: Consecutive generations –g1 and g2- evolving along the Conformational Space under selection operator. This operator selects the best individuals, enriching their statistical weight at the expenses of the diversity of the new generation.

Mathematically speaking, we are attempting to model the probability that individual, *h*, samples schema, *H*, or $P(h \in H)$. This probability of selection is proportional to both the number of instances of schema *H* in the current population, and the average fitness of schema *H* relative to the average fitness of all individuals in the current population [Eq. 9.20]:

$$P(h \in H) = \left(\frac{m(H,t)}{M}\right)\left(\frac{f(H,t)}{\bar{f}(t)}\right)$$

Equation 9.20

Where $m(H,t)$ is the number of instances of schema *H* at generation *t*, *M* is the population size, $f(H,t)$ is the mean fitness of individuals matching schema *H*, and $\bar{f}(t)$ is the mean fitness of individuals in the population.

- **Lemma 1**

Under fitness proportional selection, the expected number of instances of schema H at generation (t + 1) is [Eq. 9.21]:

$$m(H,t+1) = P(h \in H)M = m(H,t)\left(\frac{f(H,t)}{\bar{f}(t)}\right)$$

Equation 9.21

3) Search Operators: Single Point Crossover and Gene Wise Mutation

On the other hand, there is the "stochastic" force that introduces diversity and generates new chromosomes: it is represented by both the ***Mutation*** and the ***Crossover Operators***. Both of them introduce diversity in the genetic pool and enable the system to be displaced farther in the Conformational Space in the hope that better individuals will be found [Fig. 9.14].

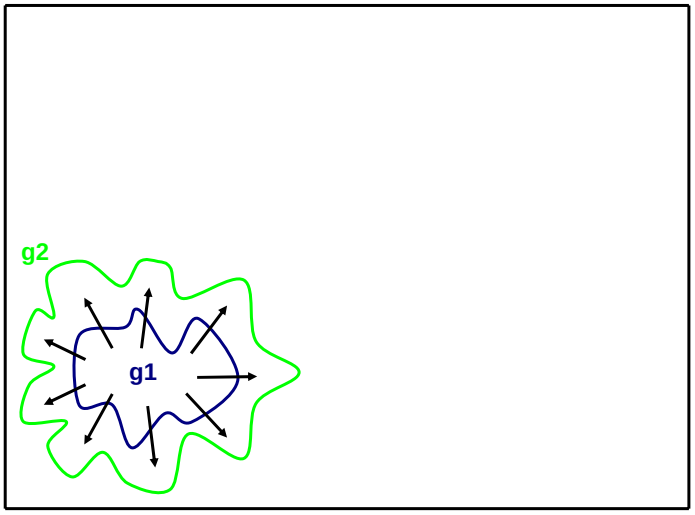


Figure 9.14: Consecutive generations –g1 and g2- evolving along Conformational Space under search operators. Both of them create randomly individuals with new schemas. The overall effect is favouring diversity and therefore the area covered by the new generations becomes increased.

Then, while the first one –Mutation- is absolutely random (except in the Canonical Genetic Algorithms), the second one –Crossover- takes into account the idea of Genetic Inheritance and let the parents to breed a new generation that carry both their genetic improvements/deteriorations.

[Single Point] ***Crossover Operator*** was the first of two search operators introduced to modify the distribution of schema in the population. Consider the following individual, *h*, two matching schema, *H1*, *H2* and crossover point between 3rd and 4th gene [Fig. 9.15]:

h	=	1	0	1	1	1	0	0
H_1	=	*	0	1	*	*	*	0
H_2	=	*	0	1	*	*	*	*

Figure 9.15

Schema *H1* will naturally be broken by the location of the crossover operator unless the second parent is able to 'repair' the disrupted gene. Nevertheless, schema *H2* emerges unaffected and is therefore independent of the second parent. Therefore, schemas with long defining lengths are more likely to be disrupted by single point crossover than schemas using short defining lengths.

- **Lemma 2**

Under single point crossover, the (lower bound) probability of schema H surviving at generation (t + 1) is, P(H survives) = 1 - P(H does not survive) [Eq. 9.22]:

$$P(H,t+1) = 1 - p_c \frac{\partial(H)}{l-1} P_{diff}(H,t)$$

Equation 9.22

Where $P_{diff}(H,t)$ is the probability that the second parent does not match schema *H*; and p_c is the a priori selected threshold of applying crossover.

[Gene Wise] ***Mutation Operator*** was the second of two search operators introduced by Holland to modify the distribution of schema in the population. Mutation is applied gene by gene and in order for schema *H* to survive, all non '*' genes in it must remain unchanged. The probability of not changing a gene is $1 - p_m$ (where p_m is the probability of mutation of a single gene) and then the survival of a given schema *H* requires that all *o(H)* non '*' genes survive [Eq. 9.23]:

$$P(H,t+1) = (1 - p_m)^{o(H)}$$

Equation 9.23

Typically, the probability of applying the mutation operator accomplishes $p_m << 1$, thus we can expand the power in the Taylor series [Eq. 9.24]:

$$(1 - p_m)^{o(H)} \approx 1 - o(H)p_m$$

Equation 9.24

- **Lemma 3**

Under gene wise mutation, the (lower bound) probability of an order o(H) schema H surviving at generation (t + 1) is [Eq. 9.25]:

$$P(H,t+1) = 1 - o(H)p_m$$

Equation 9.25

Now, combining the lemmas derived from the two opposed driving forces we obtain the final mathematical form:

- **Schema Theorem**

The expected number of schema H at generation (t + 1) when using a CGA with proportional selection, single point crossover and gene wise mutation (where the latter are applied at rates p_c and p_m) is [Eq. 9.26]:

$$m(H,t+1) \geq m(H,t)\frac{f(H,t)}{\bar{f}(t)}\left\{1 - p_c \frac{\partial(H)}{l-1} P_{diff}(H,t) - o(H)p_m\right\}$$

Equation 9.26

The theorem is described in terms of "expectation", thus strictly speaking is only true for the case of a population with an infinite number of members. The above form is specific to the selection and search operators under which it was derived. A more generic form for the Schema Theorem might take the form [Eq. 9.27] where $\alpha(H,t)$ is the "selection coefficient" and $\beta(H,t)$ is the "transcription error".

$$m(H,t+1) \geq m(H,t)\alpha(H,t)\{1 - \beta(H,t)\}$$

Equation 9.27

This is the basis for the observation that short (defining length) low order schema of above average population fitness will be favoured by CGA's. The Dynamics of this system is controlled by [Eq. 9.26] and [Eq. 9.27] and can be seen in [Fig. 9.16] as the result of the combined phenomena shown in [Fig. 9.13] and [Fig. 9.14]:

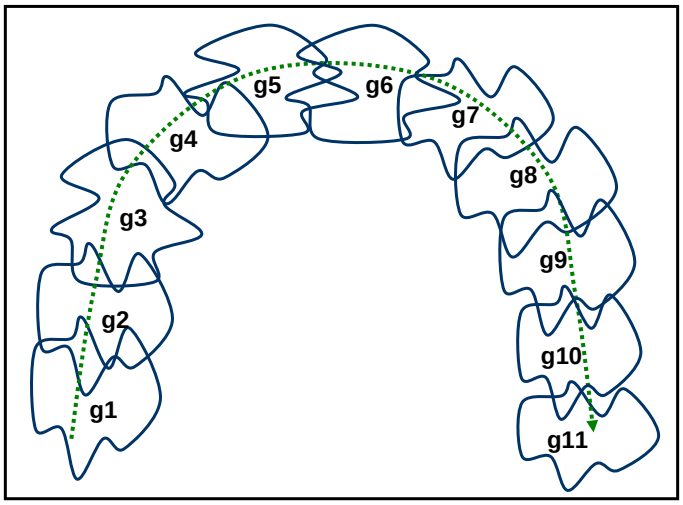


Figure 9.16: The combined effect of selection and search operators makes the system to evolve, generation after generation, over all the conformational space. The equation that rules this walk is that of Schema Theorem.

9.4.2.2.2 GACK and Macromodel

Chemistry is one of the fields of study where combinatorial optimization problems are present and therefore GA's have been already used[274]. The specific GA used in this Thesis was not the *canonical* version just explained but a variant that, in any case, behaves in a similar way, GACK (*Genetic Algorithms for Conformation Knowledge*), developed by Jonathan M. Goodman and based on a previous implementation, GACS (*Genetic Algorithms in Conformational Search*), by Nair and Goodman[275].

The program starts with a group of conformations –the first generation- randomly created from the input structure. All of them may be cut up and reassembled – crossover step- and may be mutated by having one or more of their torsion angles modified – mutation step-. The new structures are then minimised to form a new group of conformations. Then, the second generation is chosen by picking up individual conformations from both, this new group and also from the first generation. The process is iteratively repeated until the available computer time is exhausted or no new conformations are being generated [Fig. 9.17].

[274] (a) Wehrens, R.; Pretsch, E.; Buydens, L.M.C.; *J. Chem. Inf. Comput. Sci*. **1998**. *38(2)*. 151-157. (b) Meza, J.C.; Judson, R.S.; Faulkner, T.R.; Treasurywala, A.M.; *J. Comput. Chem*. **1996**. *17(9)*. 1142-1151.
[275] Nair, N.; Goodman, J.M.; *J. Chem. Inf. Comput. Sci*. **1998**. *38(2)*. 317-320.

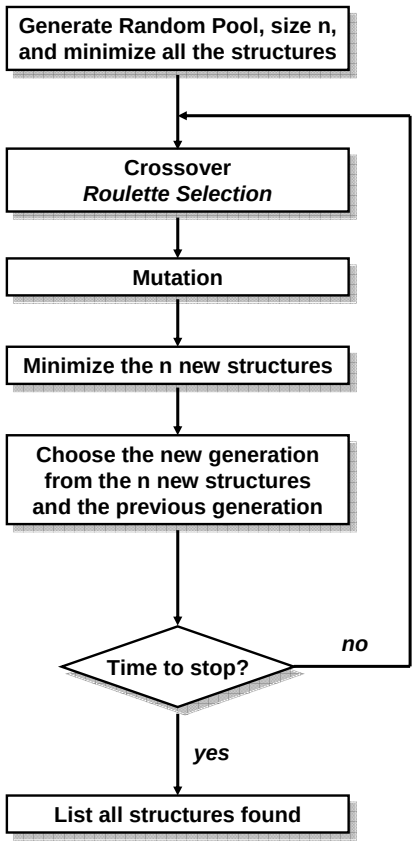


Figure 9.17: GACK's schematic chart flow.

The group of genetic operators in GACK –crossover, mutation and fitness selection- are controlled by a *roulette algorithm*[276], in which the selection probability of a particular conformation is proportional to the Boltzmann factor of its energy. These probabilities and other processes can be modulated by adjusting several operating parameters:

Simulation Parameters

- [p] Pool size.
- [g] Number of generations.
- [j] Start at generation.
- [c] Crossover rate (0.0-1.0).
- [m] Mutation rate per molecule (0.0-number of torsions).
- [s] Selection temperature (in K).
- [r] Replacement temperature (in K).
- [d] Duplication penalty (in kJ/mol).

[276] Venkatasubramanian, V.; Sundaraman, A.; *Encyclopedia of Computational Chemistry*. Vol. 2. 1115-1127. Ed. John Wiley & Sons. **1998**. ISBN 0-471-96588-X.

Operation Parameters

- [n] Pseudo random number seed.
- [x] Mutation peak width (in degrees).
- [y] Mutation peak height ratio (0.0-∞).
- [w] Crossover peak width (in degrees).
- [e] Maximum energy equivalence (in kJ/mol).
- [a] Maximum dihedral angle equivalence (in degrees).
- [f] Constraint size (in $kJ \cdot mol^{-1}q^{-2}$).
- [v] Convergence ratio fixed/relaxed.

Help

- [h] Print list of options.

The program is entirely written in Java2© and works in tandem with MacroModel 9.0[277] which handles the structural minimisation. However, the code could be developed to work with any molecular modelling program that can perform both constrained and unconstrained minimisation. GACK implements the crossover and mutation steps as a series of constraints on the molecules. The structures are minimised with these torsional constraints enforced, and then the structures are reminimised without the constraints. This usually has the effect of generating new conformations. When the calculation is finished a full ensemble is obtained, including also a conformation that is presumably the global minimum. Besides, important properties of the molecule under study can be derived from the whole group.

9.4.2.3 Simulated Annealing and Monte Carlo

Simulated Annealing (SA) is another well-known computational algorithm extensively applied to solve problems of combinatorial optimization. It was originally developed by

[277] (a) Mohamadi, F.; Richards, N.G.J.; Guida, W.C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W.C.; *J. Comput. Chem*. **1990**. *11(4)*. 440-467. (b) **MacroModel, version 9.0**; Schrödinger, LLC, New York, NY, **2005**.

Kirkpatrick, Gelatt and Vecchi[278], and is based on the Metropolis-Hastings Monte Carlo (MC) algorithm[279] and the Statistical Mechanics Theory of physical systems[280].

9.4.2.3.1 Classical Algorithm

The code is pretty similar to that of Monte Carlo, but one important difference is included when the *Cost Function* –the Energy in our case- is evaluated.

```
int i = 0;
C(i) = Cinput;
E(i) = getEnergyValue[C(i)];
while ( i <= imax && E(i) > Emax ) {
      Ctemp = perturbConformation[C(i)];
      Etemp = getEnergyValue[C(i+1)];
      if ( Etemp < E(i) ) {
            C(i+1) = Ctemp;
            E(i+1) = Etemp;
            conformationPool = store[C(i+1)];
      } else if ( P[Etemp, E(i), T(i/imax)] > random() ) {
            C(i+1) = Ctemp;
            E(i+1) = Etemp;
            conformationPool = store[C(i+1)];
      } else {
            C(i+1) = C(i);
            E(i+1) = E(i);
      }
      i = i + 1;
}
```

Both Simulated Annealing and Monte Carlo algorithms use the Maxwell-Boltzmann distribution equation as the selecting criterion [Eq. 9.28].

$$p[C(j)] = \exp\left(-\frac{E[C(j)] - E[Ctemp]}{k_B T}\right) \Rightarrow p_j = \exp\left(\frac{-\Delta E_j}{k_B T}\right)$$

Equation 9.28

Nevertheless, while Monte Carlo keeps constant Temperature all the time, Simulated Annealing changes its value along the simulation by decreasing it according to a step-dependent equation [Eq. 9.29].

[278] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.; *Science*. **1983**. *220(4598)*. 671-680.
[279] (a) Ulam, S.; Richtmyer, R.D.; von Neumann, J.; Los Alamos Scientific Laboratory Report. **1947**. *LAMS-551*. (b) Metropolis, N.; Ulam, S.; *J. Am. Statist. Ass.* **1949**. *44(247)*. 335-341.
[280] (a) Ortín, J.; Sancho, J.M.; *Curso de Física Estadística*. Ed. Universitat de Barcelona. **2001**. ISBN 84-8338-279-2. (b) Peña, M.D.; *Termodinámica Estadística*. Ed. Alhambra. Madrid. **1979**. ISBN 84-2050-661-3.

$$T(i) = (T_0 - T_{\min})\left(1 - \frac{i}{i_{\max}}\right) + T_{\min} \Rightarrow T(i) = T_0\left(1 - \frac{i}{i_{\max}}\right)$$

Equation 9.29: Equation above reduces linearly the Temperature along the SA calculation[272]. When T_{min} = 0 K is selected, the equation is easier and the simplified form on the right is derived. Other equations have been employed in further revisions.

This particular feature, along with the fact that new conformations are –generally speaking- better than the previous ones, render the algorithm stricter as the simulation reaches the last iterations. The available Conformational Space is reduced step by step forcing conformations with the best fitness ratios to be stored. As a result, the ensemble is driven to the global minimum. This process is quite similar to the one extensively used in Material Science to obtain macroscopic monocrystalline structures by means of a slow cooling under thermal equilibrium conditions, and that is why it takes its name.

9.4.2.3.2 Equivalent Thermal Algorithm

There are several SA algorithms[281]. The one used in the present Thesis is a variant of the MC more suitable for our molecules; however, it keeps the same spirit.

The classic SA includes 2 functions within the *while-loop*; firstly, we obtain a new conformation by means of *perturbConformation[C(i)]* (we were deliberately not very specific about it). And secondly, the fitness is evaluated according to the *Maxwell-Boltzmann Distribution* criterion within the *if-else* statement.

The Equivalent Thermal Algorithm is simpler and reduces the two processes –perturbation and selection- to a single one because the perturbation is internally done according to a Maxwell-Boltzmann distribution.

```
int i = 0;
C(i)= C;
while (i <= imax)
{
      C(i+1) = thermalShock[C(i)];
      conformationPool = store[C(i+1)];
      i = i + 1;
}
```

[281] (a) C.I. Chou, C.I.; Han, R.S.; Li, S.P.; Lee, T-.K.; *Phys. Rev. E*. **2003**. *67(6)*. 066704. (b) Corcho, F.J.; Filizola, M.; Pérez, J.J.; *Chem. Phys. Lett*. **2000**. *319(1-2)*. 65-70.

While the Monte Carlo provides new conformations by random *local* changes of the dihedral angles in the molecule, the thermal shock consists of a fast heating-cooling process that perturbs the *entire* macromolecule [Fig. 9.18].

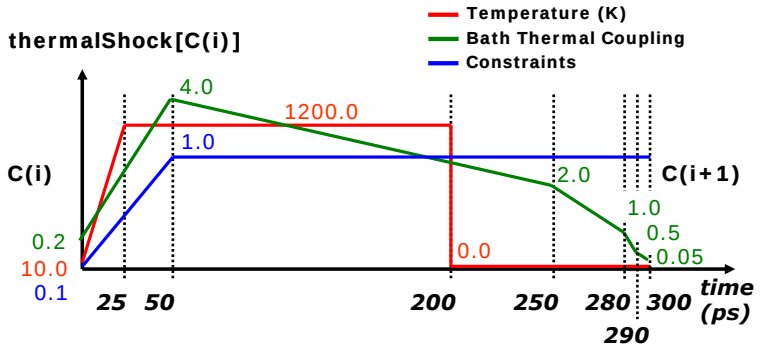


Figure 9.18: Schema of thermal Simulated Annealing. The three coloured lines represent the evolution of Temperature, bath thermal coupling, and constraints in time for a single SA step.

This is not only a large scale process involving the whole molecule; in addition it also accounts for the necessary energetic bias to reproduce a Maxwell-Boltzmann Ensemble.

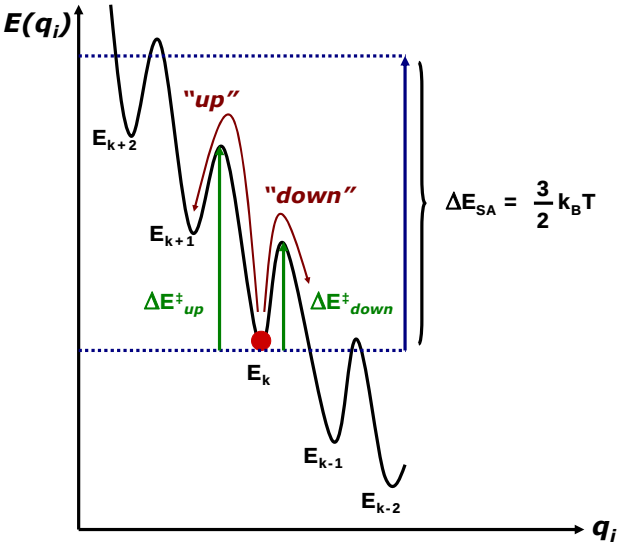


Figure 9.19: The molecule placed at minimum "k" is heated with energy enough to overpass any of the two energetic barriers; "up" and "down", however, the probability of going "down" is higher.

Given a minimum in Conformational Space, the transition probability towards a new minimum in the surrounding neighbourhood [Fig. 9.19] is proportional to the Boltzmann factor of the energetic barrier to overcome and the width of the path, ω [Eq. 9.30]:

$$p_j = \omega_j \exp\left(\frac{-\Delta E_j}{k_B T}\right)$$

Equation 9.30

The sum of all the probabilities for the total number of possible transitions in the near vicinity is a similar-type *Partition Function* [Eq. 9.31]:

$$z = \sum_n p_n = \sum_n \omega_n \exp\left(\frac{-\Delta E_n}{k_B T}\right)$$

Equation 9.31

Then, the transition probability is weighted to the probability of the total number of transitions in the surrounding Conformational Space; that is the normalised transition probability [Eq. 9.32]:

$$\rho(j) = \frac{p_j}{z} = \frac{p_j}{\sum_n p_n} = \frac{\omega_j \exp\left(\frac{-\Delta E_j}{k_B T}\right)}{\sum_n \omega_n \exp\left(\frac{-\Delta E_n}{k_B T}\right)}$$

Equation 9.32

This result provides the necessary mathematical background to prove that within a state chain of transitions, the general behaviour of the system is evolving towards better solutions, going downhill in Conformational Space [Fig. 9.20]. Hence, this behaviour is similar to that of the classical SA algorithm.

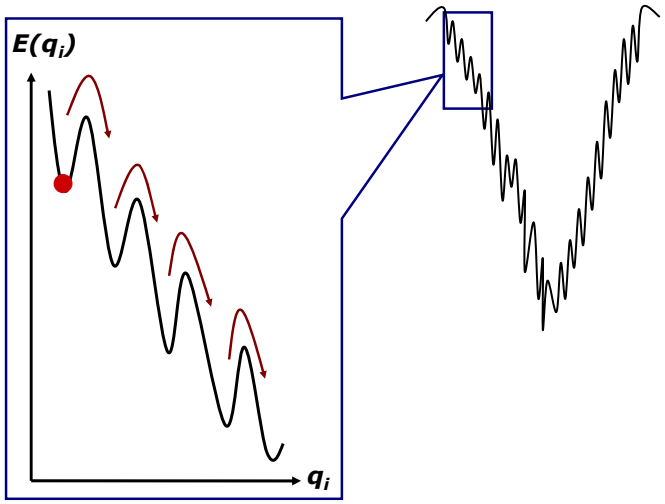


Figure 9.20: The overall behaviour of the SA thermal algorithm is that of allowing the system to descend into the global minimum area.

9.4.2.3.3 Markov Chains associated to GA, SA and MC

In fact, GA, MC and SA are *closer* than they seem. Not only they sometimes appear in mixed algorithms[282], but also –under certain conditions- the underlying physics proves that they exhibit a similar behaviour[283].

As said, MC and SA produce a conformational pool by sequentially generating new conformations –states-. Suppose we have a set of states, $S = \{s_1, s_2, s_3, \dots, s_n\}$. Upon detailed examination we realise that:

- The process starts in one of these states and moves successively from one state to another in the neighbourhood.

- If the chain is currently in state s_i, then it moves to state s_j at the next step with a transition probability denoted by p_{ij}, and this probability does not depend upon history. I.e. which states, s_n, $(n < i)$ is the chain coming from in previous steps.

Such a process, accomplishing both conditions, is said to behave as a Markov Chain, and two important properties for our research are derived:

- When the Markov Chain is long enough, the ensemble obtained follow a *Stationary Distribution* and is said to be in *Thermodynamic Equilibrium*. The necessary time/number of steps the system takes to reach this equilibrium is called *Mixing Time*[284] and the more difficult problem is to determine how many steps are needed to converge to the stationary distribution within an acceptable error.

  This is an important result in Conformational Search because it is related to the problem of how to perform an effective sampling of the Conformational Space.

[282] (a) Geyer, C.J.; Thompson, E.A.; *J. Am. Statist. Ass.* **1995**. *90(431)*.909-920. (b) Yoshida, T.; Hiroyasu, T.; Miki, M.; Ogura, M.; Okamoto, Y.; *Energy Minimization of Protein Tertiary Structure by Parallel simulated Annealing using Genetic Crossover.* Genetic and Evolutionary Computation Conference **(GECCO-2002)**, New York City, New York, USA, GECCO 2002 Workshop Program pp.49-52.
[283] ter Braak, C.J.F.; *Biometris, Wageningen UR*. April **2004**. http://www.biometris.nl
[284] Bayer, D.; Diaconis, P.; *Ann. Appl. Probab*. **1992**. *2(2)*. 294-313.

This property suggested us the idea of the *Saturation Analysis* developed in the present Thesis.

- An initial ensemble S = {s_1, s_2, s_3, … , s_n} is arranged in thermodynamic equilibrium –stationary distribution- with distribution ratios, ρ = {ρ_1, ρ_2, ρ_3, … , ρ_m}. Now, the transition probability operator is applied to every structure in the ensemble S resulting a new ensemble, S' = {s_1', s_2', s_3', … , s_n'} with distribution ratios, ρ' = {ρ_1', ρ_2', ρ_3', … , ρ_m'}. According to the Markov property, the new ensemble, S', is not only in thermodynamic equilibrium; it is also equivalent to the original ensemble S. Besides, both distribution ratios are the same, ρ = ρ'. The matrix, {π}, that converts one distribution ratio vector into the other –once the equilibrium is reached- is called *matrix of transition probabilities*, or the *transition matrix* [Eq. 9.33 to 9.35]:

$$S \Leftrightarrow S'$$
$$\rho = \rho'$$
$$\rho = \{\pi\}\rho$$

Equations 9.33; 9.34 and 9.35

This is also an important result because it indirectly provides a hint about how to evaluate whether two Conformational Search processes have been successfully done. The idea of *Analysing Overlapping Areas in the Conformational Space* for different Ensembles developed in this Thesis for MD trajectories and MC/SA was in part suggested by this property.

9.4.2.4 Molecular Dynamics

Molecular Dynamics[285] –MD- is a well-known methodology that simulates the motions of a molecule evolving in time. In this sense, it is the one that best models the behaviour of any molecule in Nature.

[285] (a) van Gunsteren, W.F.; Berendsen, H.J.C.; Rullmann, J.A.C.; *Mol. Phys.* **1981**. *44(1)*. 69-95. (b) van Gunsteren, W.F.; Berendsen, H.J.C.; *Mol. Phys.* **1982**. *45(3)*. 637-547. (c) van Gunsteren, W.F.; Berendsen, H.J.C.; *Angew. Chem. Int. Ed. Engl*. **1990**. *29(9)*. 939-1076.

It is possible to run MD "in vacuo" or in solvent just by defining a solvent box and placing your molecule right in the middle. Macroscopic variables such as temperature (T), pressure (P), number of particles (N) and volume (V) are among the most commonly used parameters that help defining the experimental conditions. It is also possible to reproduce experimental results according to the most common statistical ensembles[286] –like the Microcanonical (NVE), the Canonical (NVT) or the Isothermal-Isobaric (NPT) - specifying certain combinations of variables and switching from ones to the others.

9.4.2.4.1 Theoretical Background

The basic underlying idea in an MD methodology is supplying/removing kinetic energy to the system in a certain way so it can move all over the available potential energy surface and explore its Conformational Space [Fig. 9.21].

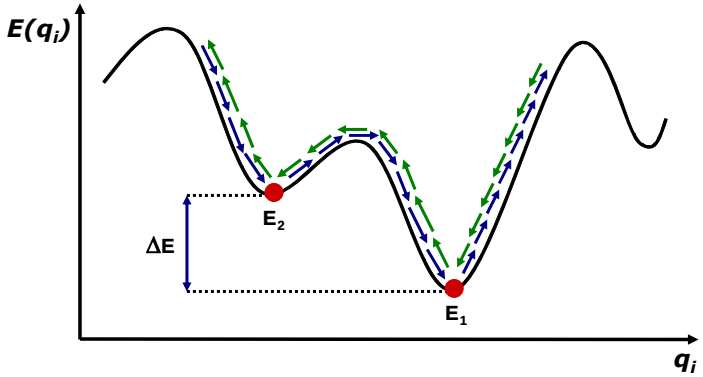


Figure 9.21: Molecule climbing barriers and moving along the potential hypersurface.

The energy control is implemented by a simulation of the system being introduced into a heat bath. Then, the heating is transferred to all the atoms assigning them random velocities compatible with the thermostatic constraint [Eq. 9.36]:

$$E = \frac{1}{2}(3N-6)k_B T = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2$$

Equation 9.36

[286] (a) Andersen, H.C.; *J. Chem. Phys*. **1980**. *72(4)*. 2384-2393. (b) Haile, J.M.; *J. Chem. Phys*. **1980**. *73(5)*. 2412-2419.

This implementation requires a pair of parameters to be specified: the *bath temperature*, T_{bath}, and also the *bath thermal coupling*, τ_T[287] . The last being an important one since it helps controlling how smoothly/tightly the system is linked to the bath [Eq. 9.37]:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}\left(T_{bath} - T(t)\right)$$

Equation 9.37

The value of τ_T, is responsible not only for the system thermal inertia but also for the type of ensemble: assuming N and V already fixed, large values of τ_T mean a weak bath-system coupling which makes it to keep total energy almost constant (NVE = Microcanonical). Nevertheless, small values lead to strong bath-system coupling and now is temperature that remains unchanged (NVT = Canonical). However, pure ensembles are rarely employed and intermediate values are often selected to allow system fluctuations to improve computational performance (they help preventing unexpected *blow up* endings by thermal shock).

A similar proceeding is also used to manage the pressure coupling by specifying the *bath pressure*, P_{bath}, and also the *bath pressure coupling*, τ_P [Eq. 9.38]:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}\left(P_{bath} - P(t)\right)$$

Equation 9.38

Pressure control is indirectly achieved by adjusting the volume of the system, which is related to pressure by means of the *isothermal compressibility*, β [Eq. 9.39].

$$\frac{dP(t)}{dt} = -\frac{1}{\beta V}\left(\frac{dV}{dt}\right)_T$$

Equation 9.39

[287] Berendsen, H.J.C.; Postma, J.P.M.; van Gunsteren, W.F.; DiNola, A.; Haak, J.R.; *J. Chem. Phys*. **1984**. *81(8)*. 3684-3690.

The velocities are scaled after a certain number of steps according to [Eq. 9.37 to 9.39]. Once the problem about how to determine the velocities is solved, the next step is integrating Newton's classical equations of motion iteratively to obtain the trajectory.

9.4.2.4.2 Algorithmic Implementation

The continuous derivatives are transformed into discrete equations under the premises of the *Finite Difference Methods*: knowing the current state of the coordinates –conformation- and the velocities –temperature- of the system, their values can be predicted at any time in the future [Fig. 9.22].

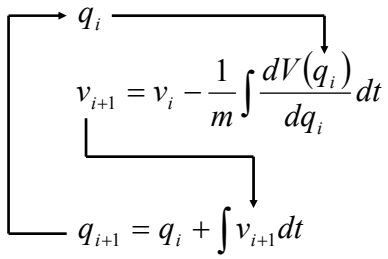


Figure 9.22: Schematic depiction of the iterative process of the integration of Newton's equations. In classical mechanics, once you determine positions and velocities at any time, all the future and past history of the system can be predicted.

There are several ways of adapting the equations of motion into iterative algorithms; the most extensively used being the *verlet*[288] and their variations –the *velocity-verlet*[289], the *leap-frog*[290] [Fig. 9.23] and the *Beeman*[291]- because of their high computational performance.

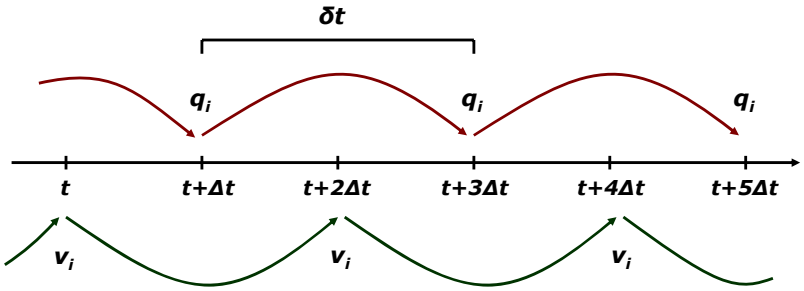


Figure 9.23: Leap-Frog algorithm. The principal feature within this algorithm is that calculates alternatively positions and momentum instead of evaluating both of them at every step. This implementation save considerably amount of time.

[288] Verlet, L.; *Phys. Rev*. **1967**. *159(1)*. 98-103.
[289] Swope, W.C.; Anderson, H.C.; Berens, P.H.; Wilson, K.R.; *J. Chem. Phys*. **1982**. *76(1)*. 637-649.
[290] Hockney, R.W.; *Meth. Comput. Phys.* **1970**. *9*. 136-211.
[291] Beeman, D.; .*J. Comput. Phys*. **1976**. *20(2)*. 130-139.

9.4.2.4.3 Sampling and Conformational Information

The output of a Molecular Dynamics calculation is not a single conformation but a trajectory, this meaning a set of conformations correlated in time.

Under detailed study, it provides valuable information: number of minima in your molecular system, size of the conformational space, conformational changes, stability, frequent conformations, folding, intra and intermolecular interactions (H-bonding, nOe), solvation energy, general energetic behaviour… All of these properties, and others, are obtained from the trajectory and their values can be regarded as trustworthy whenever the exploration of the conformational space has been effectively done. Therefore, the most important point in a MD calculation is obtaining an ensemble-representative trajectory –or at least being able to evaluate how good it is-.

In this sense, it is highly advisable to use an appropriate simulation time to ensure optimal exploration. Just as a hint, the minimum necessary time –in picoseconds- is proportional to N^2, being N the total number of atoms in your system. Under a good-enough simulation time the probabilities of visiting all the minima rise up even for the less stable, because the time ratios follow a Maxwell-Boltzmann distribution [Eq. 9.40]:

$$\frac{N_1}{N_2} = \frac{t_1}{t_2} = \frac{\rho_1}{\rho_2} = \exp\left[\frac{-\Delta E}{k_B T}\right]$$

Equation 9.40

As also said, unexpected terminations by *Blow Up* is the other problem regarding MD. There are two ways of handling this problem.

- The first one is doing the calculations in several steps, allowing the system to slowly accept the energy along the simulation. It is usually carried out dividing the MD process into, at least, three sub-steps: the first one is the *heating slope*. The second is usually the *equilibration*, which is often divided in several sub-steps –NVT and NPT- especially when running calculations involving explicit solvent molecules. And the third one is the *sampling* step –NPT- where the trajectory is actually generated and stored [Fig. 9.24].

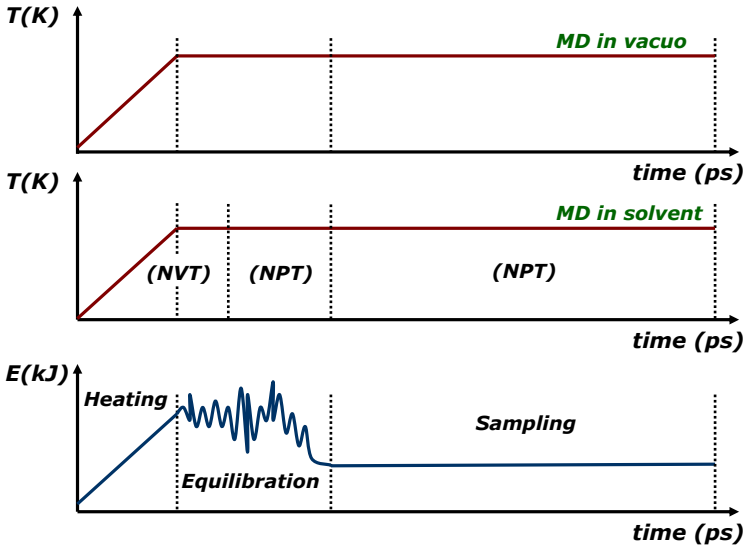


Figure 9.24: Schematic depiction of the number and type of steps in different –vacuum and solvent- MD calculations. All of them share in common the heating slope at the beginning, the central equilibration section –in one or more steps-, and the last step, where the system is sampled.

- The second one is wisely selecting the most suitable sampling frequency. In MD calculations, this parameter is represented by the *timestep* value, Δt, necessary for the numerical integrations. On the one hand, higher values lead to better explorations but also increase the probability of unexpected terminations. On the other hand, small values help avoiding *Blow Up* endings but poorly explore the Conformational Space (CS). The solution for determining the optimal timestep can be found at *Nyquist Sampling Theorem*[292] [Eq. 9.41]:

$$\Delta t = \frac{1}{2\nu_f}$$

Equation 9.41: Nyquist Theorem establishes the maximum *timestep* value necessary to sample the higher frequency motion/oscillation.

The fastest movement in a molecule is usually in the order of the stretching C-H, which is about 3000 cm^{-1}, meaning a timestep in the order of the femtosecond (10^{-15}s). This is the compromise situation between stability and CS optimal

[292] (a) Nyquist, H.; *Proc. IEEE*. **2002**. *90(2)*. 280-305. Reprint from the original paper: Nyquist, H.; *Trans. AIEE*. **1928**. *47*. 617-644. (b) Shannon, C.E.; *Proc. IEEE*. **1998**. *86(2)*. 447-457. Reprint from the original paper: Shannon, C.E.; *Proc. Institute of Radio Engineers*. **1949**. *37(1)*. 10-21. (c) Serra i Pujol, I.; Vilanova i Arbós, R.; *Tractament del Senyal*. Ed. Servei de Publicacions UAB. **1999**. ISBN 84-490-1788-2.

exploration, but still is quite a small value. SHAKE[293] algorithm, also used in this Thesis, is the answer to this problem: It is well-known that conformational changes are almost exclusively caused by torsions –low frequency motions-. This is not only the theoretical basis for the Low Mode Search Methodology[294], but also helped solving the timestep problem: High frequency molecular movements –stretching, bending…- were not only computationally expensive but also unnecessary. Therefore, by freezing any molecular movement apart from torsions, the timestep can be set to higher values. Under these conditions, better explorations of CS are obtained in reasonable simulation times.

[293] (a) Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J.C.; *J. Comput. Phys*. **1977**. *23(3)*. 327-341. (b) Andersen, H.C.; *J. Comput. Phys*. **1983**. *52(1)*. 24-34. (c) Ryckaert, J.P.; *Mol. Phys*. **1985**. *55(3)*. 549-556.
[294] Kolossváry, I.; Guida, W.C.; *J. Am. Chem. Soc*. **1996**. *118(21)*. 5011-5019.

APPENDIX (ONLY DIGITAL, NOT PRINTED)

10 APPENDIX: AMBER 7 & GAUSSIAN 98 FILES 5

10.1 SIMULATION FILES 7
10.1.1 Sander *7*
10.1.1.1 Simulated Annealing "Slow" (300 ps) –Restricted-. 7
10.1.1.2 Simulated Annealing "Fast" (30 ps) –Restricted- 7
10.1.1.3 Geometrical Optimization in vacuo –Restricted-. 8
10.1.1.4 Molecular Dynamics in vacuo 8
10.1.1.4.1 Step 1: Heating Slope and Thermal Equilibration –Restricted-. 8
10.1.1.4.2 Step 2: Sampling –Unrestricted-. 9
10.1.1.5 Molecular Dynamics in Solvent: Water 9
10.1.1.5.1 Step 0: Geometrical Optimization –Restricted-. 9
10.1.1.5.2 Step 1: Heating Slope and Thermal Equilibration (NVT) –Restricted-. 9
10.1.1.5.3 Step 2: Thermal Equilibration (NPT) –Unrestricted-. 10
10.1.1.5.4 Step 3: Sampling (NPT) –Unrestricted-. 10
10.1.1.6 Molecular Dynamics in Solvent: Benzene 10
10.1.1.6.1 Step 0: Geometrical Optimization –Restricted-. 10
10.1.1.6.2 Step 1: Heating Slope and Thermal Equilibration (NVT) –Restricted-. 11
10.1.1.6.3 Step 2: Thermal Equilibration under (NPT) –Unrestricted-. 11
10.1.1.6.4 Step 3: Sampling under (NPT) –Unrestricted-. 12
10.1.1.7 Constraints. 12
10.2 ANALYSIS FILES 13
10.2.1 Ptraj *13*
10.2.1.1 Dihedral Angle O3(n)-C4(n)-C1(n+1)-O2(n+1) 13
10.2.1.2 Hydrogen Bonding. 13
10.2.1.2.1 In vacuo. 13
10.2.1.2.2 In Solvent H_2O. 13
10.2.2 Carnal *13*
10.2.2.1 Hydrogen Bonding 13
10.2.2.2 Radius of Gyration 14
10.2.2.3 RMSD Conformations 14
10.3 MISCELLANEA 14
10.3.1 Ptraj *14*
10.3.1.1 Extracting Solvent 14
10.3.1.2 Average Structure 14
10.3.1.3 Packing a full set of individual structures into a single file 14
10.3.1.4 Trajectory transformation: AMBER (.trj) to CHARMM (.dcd) 15
10.3.1.4.1 In vacuo 15
10.3.1.4.2 In H_2O. 15
10.4 GAUSSIAN 98 FILES 15
10.4.1 Geometrical Optimization / Energy Minimization *15*
10.4.2 Merz-Kollman Charges Calculation *15*

11 APPENDIX: SET OF CHARGES 17

11.1 INTRODUCTION 19
11.2 GLUCOSE: UNITS FOR AMBER 7 19
11.2.1 Charges by Dr. Iván Beà *19*
11.2.2 Charges by Dr. Javier Pérez Mirón *20*
11.3 BENZENE SOLVENT MODEL: UNITS FOR AMBER 7 20
11.3.1 The Set of Charges *21*
11.3.2 Conclusions *23*
11.3.3 Redefining the Problem *23*
11.3.3.1 The Electrostatic Approach 23
11.3.3.2 The van der Waals Approach 25

12 APPENDIX: C-SHELL SCRIPTS AND JAVA2 SOFTWARE 27

12.1 NAMEGIANTCYD.JAVA [JAVA2]: IMPROVED AUTOMATIC NOMENCLATURE 29
12.2 SORTDIHEDRAL.JAVA [JAVA2]: REMOVES CDS CYCLIC SYMMETRY BEFORE PCA 47
12.3 MARKOVCHAINANALYSIS.JAVA [JAVA2]: NOMENCLATURE STATISTICS AND MARKOV TRANSITION MATRIX 60
12.4 STATISTICS.EXE [C-SHELL SCRIPT]. SATURATION ANALYSIS FOR SIMULATED ANNEALING AND MOLECULAR DYNAMICS 76

12.5 DYNAMICLAMBDA.EXE [C-SHELL SCRIPT]. DYNAMIC LAMBDA INDEX FOR MOLECULAR DYNAMICS 79

10 APPENDIX: AMBER 7 & GAUSSIAN 98 FILES

10.1 Simulation Files

10.1.1 Sander

10.1.1.1 Simulated Annealing "Slow" (300 ps) –Restricted-.

```
Simulated Annealing calculation for ca05SA - Slow Calculation.
 &cntrl
imin=0, nmropt=1,
ntx=1, irest=0, ntrx=1,
ntxo=1, ntpr=1000, ntwr=1000,
iwrap=0, ntwx=1000, ntwv=0, ntwe=1000,
lastrst=3000000,
ioutfm=0, ntwprt=0, idecomp=0,
ntf=2, ntb=0, igb=0, nsnb=25,
ipol=0, gbsa=0,
dielc=1.0, cut=12.0, intdiel=1.0,
scnb=2.0, scee=1.2,
nstlim=300000, nscm=1000, nrespa=1,
t=0.0, dt=0.001, vlimit=10.0,
ig=71277, ntt=1, vrand=0,
temp0= 1000.0, tempi=0.0, heat=0.0,
dtemp=5.0, tautp=0.5,
ntc=2,
tol=0.00001
 &end
 &wt type='TEMP0', istep1=1,      istep2=25000,  value1=10.0,   value2=1200.0, &end
 &wt type='TEMP0', istep1=25001,  istep2=200000, value1=1200.0, value2=1200.0, &end
 &wt type='TEMP0', istep1=200001, istep2=300000, value1=0.0,    value2=0.0,    &end
 &wt type='TAUTP', istep1=1,      istep2=50000,  value1=0.2,    value2=4.0,    &end
 &wt type='TAUTP', istep1=50001,  istep2=250000, value1=4.0,    value2=2.0,    &end
 &wt type='TAUTP', istep1=250001, istep2=280000, value1=2.0,    value2=1.0,    &end
 &wt type='TAUTP', istep1=280001, istep2=290000, value1=1.0,    value2=0.5,    &end
 &wt type='TAUTP', istep1=290001, istep2=300000, value1=0.5,    value2=0.05,   &end
 &wt type='REST',  istep1=1,      istep2=50000,  value1=0.1,    value2=1.0,    &end
 &wt type='REST',  istep1=50001,  istep2=300000, value1=1.0,    value2=1.0,    &end
 &wt type='END'
 &end
DISANG=ca05SA rst.f
```

10.1.1.2 Simulated Annealing "Fast" (30 ps) –Restricted-.

```
Simulated Annealing calculation for ca14SA - Fast Calculation
 &cntrl
imin=0, nmropt=1,
ntx=1, irest=0, ntrx=1,
ntxo=1, ntpr=1000, ntwr=1000,
iwrap=0, ntwx=1000, ntwv=0, ntwe=1000,
lastrst=3000000,
ioutfm=0, ntwprt=0, idecomp=0,
ntf=2, ntb=0, igb=0, nsnb=25,
ipol=0, gbsa=0,
dielc=1.0, cut=12.0, intdiel=1.0,
scnb=2.0, scee=1.2,
nstlim=30000, nscm=1000, nrespa=1,
t=0.0, dt=0.001, vlimit=10.0,
ig=71277, ntt=1, vrand=0,
temp0= 1000.0, tempi=0.0, heat=0.0,
dtemp=5.0, tautp=0.5,
ntc=2,
tol=0.00001
 &end
 &wt type='TEMP0', istep1=1,     istep2=2500,  value1=10.0,   value2=1200.0, &end
 &wt type='TEMP0', istep1=2501,  istep2=20000, value1=1200.0, value2=1200.0, &end
```

```
 &wt type='TEMP0', istep1=20001, istep2=30000, value1=0.0,    value2=0.0,    &end
 &wt type='TAUTP', istep1=1,     istep2=5000,  value1=0.2,    value2=4.0,    &end
 &wt type='TAUTP', istep1=5001,  istep2=25000, value1=4.0,    value2=2.0,    &end
 &wt type='TAUTP', istep1=25001, istep2=28000, value1=2.0,    value2=1.0,    &end
 &wt type='TAUTP', istep1=28001, istep2=29000, value1=1.0,    value2=0.5,    &end
 &wt type='TAUTP', istep1=29001, istep2=30000, value1=0.5,    value2=0.05,   &end
 &wt type='REST',  istep1=1,     istep2=5000,  value1=0.1,    value2=1.0,    &end
 &wt type='REST',  istep1=5001,  istep2=30000, value1=1.0,    value2=1.0,    &end
 &wt type='END'
 &end
DISANG=ca14SA_rst.f
```

10.1.1.3 Geometrical Optimization in vacuo –Restricted-.

```
Geometrical Optimization. No box. Restraints.
 &cntrl
imin=1, nmropt=1,
ntx=1, irest=0, ntrx=1,
ntb=0, igb=0, nsnb=25,
ipol=0, gbsa=0,
dielc=1.0, cut=12.0, intdiel=1.0,
scnb=2.0, scee=1.2,
ibelly=0, ntr=0,
maxcyc=5000,
 &end
1.0
 &wt type='END' &end
DISANG=ca14SA_rst.f
```

10.1.1.4 Molecular Dynamics in vacuo

10.1.1.4.1 Step 1: Heating Slope and Thermal Equilibration –Restricted-.

```
MD in vacuo. Heating and Thermal Equilibration.
 &cntrl
imin=0, nmropt=1,
ntx=1, irest=0, ntrx=1,
ntxo=1, ntpr=1000, ntwr=1000,
iwrap=0, ntwx=1000, ntwv=0, ntwe=1000,
lastrst=3000000,
ioutfm=0, ntwprt=0, idecomp=0,
ntf=2, ntb=0, igb=0, nsnb=25,
ipol=0, gbsa=0,
dielc=1.0, cut=12.0, intdiel=1.0,
scnb=2.0, scee=1.2,
nstlim=300000, nscm=1000, nrespa=1,
t=0.0, dt=0.001, vlimit=10.0,
ig=71277, ntt=1, vrand=0,
temp0=298.0, tempi=0.0, heat=0.0,
dtemp=5.0, tautp=0.5,
ntc=2,
tol=0.00001
 &end
 &wt type='TEMP0', istep1=1,      istep2=100000, value1=10.0,   value2=298.0, &end
 &wt type='TEMP0', istep1=100001, istep2=300000, value1=298.0,  value2=298.0, &end
 &wt type='TAUTP', istep1=1,      istep2=50000,  value1=0.2,    value2=4.0,   &end
 &wt type='TAUTP', istep1=50001,  istep2=150000, value1=4.0,    value2=2.0,   &end
 &wt type='TAUTP', istep1=150001, istep2=200000, value1=2.0,    value2=0.5,   &end
 &wt type='TAUTP', istep1=200001, istep2=300000, value1=0.5,    value2=0.5,   &end
 &wt type='REST',  istep1=1,      istep2=50000,  value1=0.1,    value2=1.0,   &end
 &wt type='REST',  istep1=50001,  istep2=250000, value1=1.0,    value2=1.0,   &end
 &wt type='REST',  istep1=250001, istep2=300000, value1=1.0,    value2=0.1,   &end
 &wt type='END'
 &end
DISANG=ca05DM_rst.f
```

10.1.1.4.2 Step 2: Sampling –Unrestricted-.

```
MD in vacuo. Sampling step.
 &cntrl
imin=0, nmropt=0,
ntx=5, irest=1, ntrx=1,
ntxo=1, ntpr=10000, ntwr=10000,
iwrap=0, ntwx=10000, ntwv=0, ntwe=10000,
lastrst=3000000,
ioutfm=0, ntwprt=0, idecomp=0,
ntf=2, ntb=0, igb=0, nsnb=25,
ipol=0, gbsa=0,
dielc=1.0, cut=12.0, intdiel=1.0,
scnb=2.0, scee=1.2,
nstlim=10000000, nscm=1000, nrespa=1,
t=0.0, dt=0.001, vlimit=10.0,
ig=71277, ntt=1, vrand=0,
temp0=298.0, tempi=298.0, heat=0.0,
dtemp=5.0, tautp=0.5,
ntc=2,
tol=0.00001
 &end
```

10.1.1.5 Molecular Dynamics in Solvent: Water

10.1.1.5.1 Step 0: Geometrical Optimization –Restricted-.

```
stp0. Geometrical Optimization. Restrictions.
 &cntrl
   imin=1,     nmropt=1,
   ntx=1,      irest=0,     ntrx=1,
   ntb=1,      igb=0,       nsnb=25,
   ipol=0,     gbsa=0,
   dielc=1.0,  cut=12.0,    intdiel=1.0,
   scnb=2.0,   scee=1.2,
   ibelly=0,   ntr=0,
   maxcyc=50000,
 &end
1.0
 &wt type='END' &end
DISANG=ca08i_rst.f
```

10.1.1.5.2 Step 1: Heating Slope and Thermal Equilibration (NVT) – Restricted-.

```
stp1. Heating Slope and Thermal Equilibration. Constant Volume (NVT). Restrictions.
 &cntrl
   imin=0,         nmropt=1,
   ntx=1,          irest=0,     ntrx=1,
   ntxo=1,         ntpr=1000,   ntwr=1000,
   iwrap=1,
   ntwx=1000,      ntwv=0,      ntwe=1000,
   lastrst=50000000,
   ntf=2,          ntb=1,
   igb=0,          nsnb=25,
   ntc=2,          cut=12.0,
   nstlim=300000,  nscm=1000,   nrespa=1,
   t=0.0,          dt=0.001,    vlimit=20.0,
   ig=71277,       ntt=1,       vrand=0,
   temp0=298.0,    tempi=0.0,   heat=0.0,
   dtemp=5.0,      tautp=0.5,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0,      istep2=100000, value1=0.0,   value2=298.0, &end
 &wt type='TEMP0', istep1=100001, istep2=300000, value1=298.0, value2=298.0, &end
```

```
 &wt type='TAUTP', istep1=1,      istep2=50000,  value1=0.2,   value2=4.0,   &end
 &wt type='TAUTP', istep1=50001,  istep2=250000, value1=4.0,   value2=2.0,   &end
 &wt type='TAUTP', istep1=250001, istep2=280000, value1=2.0,   value2=1.0,   &end
 &wt type='TAUTP', istep1=280001, istep2=290000, value1=1.0,   value2=0.5,   &end
 &wt type='TAUTP', istep1=290001, istep2=300000, value1=0.5,   value2=0.05,  &end
 &wt type='REST',  istep1=1,      istep2=100000, value1=0.1,   value2=1.0,   &end
 &wt type='REST',  istep1=100001, istep2=300000, value1=1.0,   value2=1.0,   &end
 &wt type='END'
 &end
DISANG=ca08i_rst.f
```

10.1.1.5.3 Step 2: Thermal Equilibration (NPT) –Unrestricted-.

```
stp2. Thermal Equilibration. Constant Pressure (NPT). No Restrictions.
 &cntrl
   imin=0,          nmropt=1,
   ntx=1,           irest=0,      ntrx=1,
   ntxo=1,          ntpr=1000,    ntwr=1000,
   iwrap=1,
   ntwx=1000,       ntwv=0,       ntwe=1000,
   lastrst=50000000,
   ntf=2,           ntb=2,        ntp=1,
   igb=0,           nsnb=25,
   ntc=2,           cut=12.0,
   nstlim=300000,   nscm=1000,    nrespa=1,
   t=0.0,           dt=0.001,     vlimit=20.0,
   ig=71277,        ntt=1,        vrand=0,
   temp0=298.0,     tempi=298.0, heat=0.0,
   dtemp=5.0,       tautp=0.5,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0, istep2=300000, value1=298.0, value2=298.0, &end
 &wt type='END'
 &end
```

10.1.1.5.4 Step 3: Sampling (NPT) –Unrestricted-.

```
stp3. Sampling (NPT). No Restrictions.
 &cntrl
   imin=0,          nmropt=1,
   ntx=1,           irest=0,      ntrx=1,
   ntxo=1,          ntpr=1000,    ntwr=1000,
   iwrap=0,
   ntwx=1000,       ntwv=0,       ntwe=1000,
   lastrst=50000000,
   ntf=2,           ntb=2,        ntp=1,
   igb=0,           nsnb=25,
   ntc=2,           cut=12.0,
   nstlim=5000000, nscm=1000,     nrespa=1,
   t=0.0,           dt=0.001,     vlimit=20.0,
   ig=71277,        ntt=1,        vrand=0,
   temp0=298.0,     tempi=298.0,  heat=0.0,
   dtemp=5.0,       tautp=0.5,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0, istep2=5000000, value1=298.0, value2=298.0, &end
 &wt type='END'
 &end
```

10.1.1.6 Molecular Dynamics in Solvent: Benzene

10.1.1.6.1 Step 0: Geometrical Optimization –Restricted-.

```
stp0. Geometrical Optimization. Restrictions.
 &cntrl
```

```
   imin=1,     nmropt=1,
   ntx=1,      irest=0,     ntrx=1,
   ntb=1,      igb=0,       nsnb=25,
   ipol=0,     gbsa=0,
   dielc=1.0,  cut=12.0,    intdiel=1.0,
   scnb=2.0,   scee=1.2,
   ibelly=0,   ntr=0,
   maxcyc=50000,
 &end
1.0
 &wt type='END' &end
DISANG=ca26i_rst.f
```

10.1.1.6.2 Step 1: Heating Slope and Thermal Equilibration (NVT) – Restricted-.

```
stp1. Heating Slope and Thermal Equilibration. Constant Volume (NVT). Restrictions.
 &cntrl
   imin=0,          nmropt=1,
   ntx=1,           irest=0,     ntrx=1,
   ntxo=1,          ntpr=1000,   ntwr=1000,
   iwrap=1,
   ntwx=1000,       ntwv=0,      ntwe=1000,
   lastrst=50000000,
   ntf=2,           ntb=1,
   igb=0,           nsnb=25,
   ntc=2,           cut=12.0,
   nstlim=300000,   nscm=1000,   nrespa=1,
   t=0.0,           dt=0.001,    vlimit=20.0,
   ig=71277,        ntt=1,       vrand=0,
   temp0=298.0,     tempi=0.0,   heat=0.0,
   dtemp=5.0,       tautp=0.5,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0,      istep2=100000, value1=0.0,   value2=298.0, &end
 &wt type='TEMP0', istep1=100001, istep2=300000, value1=298.0, value2=298.0, &end
 &wt type='TAUTP', istep1=1,      istep2=50000,  value1=0.2,   value2=4.0,   &end
 &wt type='TAUTP', istep1=50001,  istep2=250000, value1=4.0,   value2=2.0,   &end
 &wt type='TAUTP', istep1=250001, istep2=280000, value1=2.0,   value2=1.0,   &end
 &wt type='TAUTP', istep1=280001, istep2=290000, value1=1.0,   value2=0.5,   &end
 &wt type='TAUTP', istep1=290001, istep2=300000, value1=0.5,   value2=0.05,  &end
 &wt type='REST',  istep1=1,      istep2=100000, value1=0.1,   value2=1.0,   &end
 &wt type='REST',  istep1=100001, istep2=300000, value1=1.0,   value2=1.0,   &end
 &wt type='END'
 &end
DISANG=ca26i_rst.f
```

10.1.1.6.3 Step 2: Thermal Equilibration under (NPT) –Unrestricted-.

```
stp2. Thermal Equilibration. Constant Pressure (NPT). No Restrictions.
 &cntrl
   imin=0,          nmropt=1,
   ntx=1,           irest=0,     ntrx=1,
   ntxo=1,          ntpr=1000,   ntwr=1000,
   iwrap=1,
   ntwx=1000,       ntwv=0,      ntwe=1000,
   lastrst=50000000,
   ntf=2,           ntb=2,       ntp=1,
   igb=0,           nsnb=25,
   ntc=2,           cut=12.0,
   nstlim=300000,   nscm=1000,   nrespa=1,
   t=0.0,           dt=0.001,    vlimit=20.0,
   ig=71277,        ntt=1,       vrand=0,
   temp0=298.0,     tempi=298.0, heat=0.0,
   dtemp=5.0,       tautp=0.5,   comp=96.7,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0, istep2=300000, value1=298.0, value2=298.0, &end
 &wt type='END'
 &end
```

10.1.1.6.4 Step 3: Sampling under (NPT) –Unrestricted-.

```
stp3. Sampling (NPT). No Restrictions.
 &cntrl
   imin=0,          nmropt=1,
   ntx=1,           irest=0,       ntrx=1,
   ntxo=1,          ntpr=1000,     ntwr=1000,
   iwrap=0,
   ntwx=1000,       ntwv=0,        ntwe=1000,
   lastrst=50000000,
   ntf=2,           ntb=2,         ntp=1,
   igb=0,           nsnb=25,
   ntc=2,           cut=12.0,
   nstlim=5000000, nscm=1000,     nrespa=1,
   t=0.0,           dt=0.001,      vlimit=20.0,
   ig=71277,        ntt=1,         vrand=0,
   temp0=298.0,     tempi=298.0,   heat=0.0,
   dtemp=5.0,       tautp=0.5,     comp=96.7,
   tol=0.00001,
 &end
 &wt type='TEMP0', istep1=0, istep2=5000000, value1=298.0, value2=298.0, &end
 &wt type='END'
 &end
```

10.1.1.7 Constraints.

ca05SA_rst.f

```
&rst iat= 1, 11, 12, 19, r1=27, r2=55, r3=59, r4=87, rk2=500.0, rk3=500.0, &end
&rst iat= 3, 1, 11, 12, r1=-88, r2=-60, r3=-56, r4=-28, rk2=500.0, rk3=500.0, &end
&rst iat= 7, 3, 1, 11, r1=25, r2=53, r3=57, r4=85, rk2=500.0, rk3=500.0, &end
&rst iat= 11, 12, 19, 7, r1=-80, r2=-52, r3=-48, r4=-20, rk2=500.0, rk3=500.0, &end
&rst iat= 12, 19, 7, 3, r1=21, r2=49, r3=53, r4=81, rk2=500.0, rk3=500.0, &end
&rst iat= 19, 7, 3, 1, r1=-84, r2=-56, r3=-52, r4=-24, rk2=500.0, rk3=500.0, &end
&rst iat= 22, 32, 33, 40, r1=23, r2=51, r3=55, r4=83, rk2=500.0, rk3=500.0, &end
&rst iat= 24, 22, 32, 33, r1=-91, r2=-63, r3=-59, r4=-31, rk2=500.0, rk3=500.0, &end
&rst iat= 28, 24, 22, 32, r1=30, r2=58, r3=62, r4=90, rk2=500.0, rk3=500.0, &end
&rst iat= 32, 33, 40, 28, r1=-73, r2=-45, r3=-41, r4=-13, rk2=500.0, rk3=500.0, &end
&rst iat= 33, 40, 28, 24, r1=16, r2=44, r3=48, r4=76, rk2=500.0, rk3=500.0, &end
&rst iat= 40, 28, 24, 22, r1=-85, r2=-57, r3=-53, r4=-25, rk2=500.0, rk3=500.0, &end
&rst iat= 43, 53, 54, 61, r1=23, r2=51, r3=55, r4=83, rk2=500.0, rk3=500.0, &end
&rst iat= 45, 43, 53, 54, r1=-96, r2=-68, r3=-64, r4=-36, rk2=500.0, rk3=500.0, &end
&rst iat= 49, 45, 43, 53, r1=34, r2=62, r3=66, r4=94, rk2=500.0, rk3=500.0, &end
&rst iat= 53, 54, 61, 49, r1=-66, r2=-38, r3=-34, r4=-6, rk2=500.0, rk3=500.0, &end
&rst iat= 54, 61, 49, 45, r1=8, r2=36, r3=40, r4=68, rk2=500.0, rk3=500.0, &end
&rst iat= 61, 49, 45, 43, r1=-82, r2=-54, r3=-50, r4=-22, rk2=500.0, rk3=500.0, &end
&rst iat= 64, 74, 75, 82, r1=27, r2=55, r3=59, r4=87, rk2=500.0, rk3=500.0, &end
&rst iat= 66, 64, 74, 75, r1=-91, r2=-63, r3=-59, r4=-31, rk2=500.0, rk3=500.0, &end
&rst iat= 70, 66, 64, 74, r1=29, r2=57, r3=61, r4=89, rk2=500.0, rk3=500.0, &end
&rst iat= 74, 75, 82, 70, r1=-78, r2=-50, r3=-46, r4=-18, rk2=500.0, rk3=500.0, &end
&rst iat= 75, 82, 70, 66, r1=19, r2=47, r3=51, r4=79, rk2=500.0, rk3=500.0, &end
&rst iat= 82, 70, 66, 64, r1=-85, r2=-57, r3=-53, r4=-25, rk2=500.0, rk3=500.0, &end
&rst iat= 85, 95, 96, 103, r1=20, r2=48, r3=52, r4=80, rk2=500.0, rk3=500.0, &end
&rst iat= 87, 85, 95, 96, r1=-93, r2=-65, r3=-61, r4=-33, rk2=500.0, rk3=500.0, &end
&rst iat= 91, 87, 85, 95, r1=31, r2=59, r3=63, r4=91, rk2=500.0, rk3=500.0, &end
&rst iat= 95, 96, 103, 91, r1=-66, r2=-38, r3=-34, r4=-6, rk2=500.0, rk3=500.0, &end
&rst iat= 96, 103, 91, 87, r1=9, r2=37, r3=41, r4=69, rk2=500.0, rk3=500.0, &end
&rst iat= 103, 91, 87, 85, r1=-83, r2=-55, r3=-51, r4=-23, rk2=500.0, rk3=500.0, &end
```

10.2 Analysis Files

10.2.1 Ptraj

10.2.1.1 Dihedral Angle O3(n)-C4(n)-C1(n+1)-O2(n+1)

```
trajin  /disk9/segurKepa/iter05.abgCyD_SA/ca05/ca05SA1.rst
dihedral  ang1  :1@O3  :1@C4  :2@C1  :2@O2 out  O3C4C1O2_1.txt
dihedral  ang2  :2@O3  :2@C4  :3@C1  :3@O2 out  O3C4C1O2_2.txt
dihedral  ang3  :3@O3  :3@C4  :4@C1  :4@O2 out  O3C4C1O2_3.txt
dihedral  ang4  :4@O3  :4@C4  :5@C1  :5@O2 out  O3C4C1O2_4.txt
dihedral  ang5  :5@O3  :5@C4  :1@C1  :1@O2 out  O3C4C1O2_5.txt
go
```

10.2.1.2 Hydrogen Bonding.

10.2.1.2.1 In vacuo.

```
trajin  /disk9/kepa/aca/ca05SA_full.trj.gz
donor    mask :*@O?
acceptor mask :*@O2 :*@H2O
acceptor mask :*@O3 :*@H3O
acceptor mask :*@O6 :*@H6O
hbond distance 3.5 includeself
go
```

10.2.1.2.2 In Solvent H_2O.

```
trajin  /disk9/kepa/aka/ca08ih2os1dm_stp3.trj.gz
donor    mask :*@O?
acceptor mask :*@O2 :*@H2O
acceptor mask :*@O3 :*@H3O
acceptor mask :*@O6 :*@H6O
hbond distance 3.5 solventneighbor 6 solventdonor WAT O solventacceptor WAT O H1
solventacceptor WAT O H2 includeself
go
```

10.2.2 Carnal

10.2.2.1 Hydrogen Bonding

```
FILES_IN
 PARM   p1 /disk9/kepa/aka/1024SAitzi/ca05/ca05SA.top;
 STREAM s1 /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.trj;
FILES_OUT
 HBOND h1  /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.hbd LIST;
DECLARE
 GROUP grp1 ( ATOM TYPE OH );
 GROUP grp2 ( ATOM TYPE OH OS );
OUTPUT
 HBOND h1 DONOR grp1 ACCEPTOR grp2 STATS;
END
```

10.2.2.2 Radius of Gyration

```
FILES_IN
 PARM   p1  /disk9/kepa/aka/1024SAitzi/ca05/ca05SA.top;
 STREAM s1  /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.trj;
FILES_OUT
 TABLE tab1 /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.gyr;
DECLARE
 GROUP grp1 ( ATOM NAME C? O? H? );
OUTPUT
 TABLE tab1 grp1%radgyr;
END
```

10.2.2.3 RMSD Conformations

```
FILES_IN
 PARM   p1      /disk9/kepa/aka/1024SAitzi/ca05/ca05SA.top;
 STREAM s1      /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.trj;
 STATIC refConf /disk9/kepa/aka/1024SAitzi/ca05/ca05SA1025.rst;
FILES_OUT
 TABLE tab1     /disk9/kepa/aka/1024SAitzi/ca05/ca05SA_full.rms;
DECLARE
 GROUP grp1 ( ATOM NAME C? O? H? );
 RMS   fit1 FIT grp1 s1 refConf;
OUTPUT
 TABLE tab1 fit1;
END
```

10.3 Miscellanea

10.3.1 Ptraj

10.3.1.1 Extracting Solvent

```
trajin  /disk9/kepa/iter12.abgCyD_MD/ca26iBNZ/conf01/ca26iBNZs1DM_stp3.trj
strip :BNZ
center
trajout /disk9/kepa/iter12.abgCyD_MD/ca26iBNZ/conf01/ca26iBNZs1DM_stp3_noSOLV.trj
go
```

10.3.1.2 Average Structure

```
trajin  /disk9/kepa/aka/ca08iH2Os1DM_stp3_noSOLV.trj.gz
rms previous
average /disk9/kepa/aka/ca08iH2Os1DM.raw_AVE.rst nobox restart
go
```

10.3.1.3 Packing a full set of individual structures into a single file

```
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA1.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA2.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA3.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA4.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA5.rst  restrt

… (and so on until we reach the last lines) …
```

```
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA1022.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA1023.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA1024.rst  restrt
trajin  /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA1025.rst  restrt
center
rms previous
trajout  /disk7/kepa/iter05.abgCyD SA/ca05/ca05SA full.trj  trajectory nobox
```

10.3.1.4 Trajectory transformation: AMBER (.trj) to CHARMM (.dcd)

10.3.1.4.1 In vacuo

```
trajin /disk7/kepa/iter05.abgCyD_SA/ca05/ca05SA_full.trj
rms previous
trajout ca05SA_full.dcd charmm nobox little
go
```

10.3.1.4.2 In H_2O.

```
trajin  /disk8/kepa/ca08iH2O/conf01/ca08iH2Os1DM_stp3.trj.gz
strip :WAT
center
trajout /disk8/kepa/ca08iH2O/conf01/ca08iH2Os1DM_stp3_noSOLV.dcd charmm nobox little
go
```

10.4 Gaussian 98 Files

10.4.1 Geometrical Optimization / Energy Minimization

```
%nproc=2
$Mem=500Mb
%chk=/cescascratch/uabqur01/kepa/bnz.chk
#p HF/STO-3G guess=read Opt=(restart) scf=(save,maxcyc=500) pop=none
geom(nodistance,noangle)

Optimization Benzene

0 1

Z-MATRIX
```

10.4.2 Merz-Kollman Charges Calculation

```
%nproc=2
$Mem=500Mb
%chk=/cescascratch/uabqur01/kepa/bnz.chk
#N HF/6-31G* pop=mk iop(6/33=2) guess=read geom(checkpoint,nodistance,noangle)

Merz Kollman Charges Benzene

0 1
```

11 APPENDIX: SET OF CHARGES

11.1 Introduction

This chapter contains the set of charges employed in the present Thesis. Those for glucose unit were already calculated by former member of our group. The charges for the solvent benzene unit were developed in this study.

11.2 Glucose: Units for AMBER 7

The two different set of charges for glucose are shown hereafter.

11.2.1 Charges by Dr. Iván Beà

The set was calculated using the same conditions employed to parameterise the parm99 Force Field. In the first place, the structure was created and minimised using Molecular Mechanics with Macromodel. On the second place, the structure was minimised using Quantum Mechanics at *Ab Initio* level with Gaussian 98: Methodology Hartree Fock with STO-3G basis set and a maximum of 500 steps was employed. Finally, *Ab Initio* Hartree Fock level with a 6-31G* basis set and Merz-Singh-Kollman charges was used. The Gaussian output file was processed with RESP module to obtain the final set of charges [Fig. 11.1].

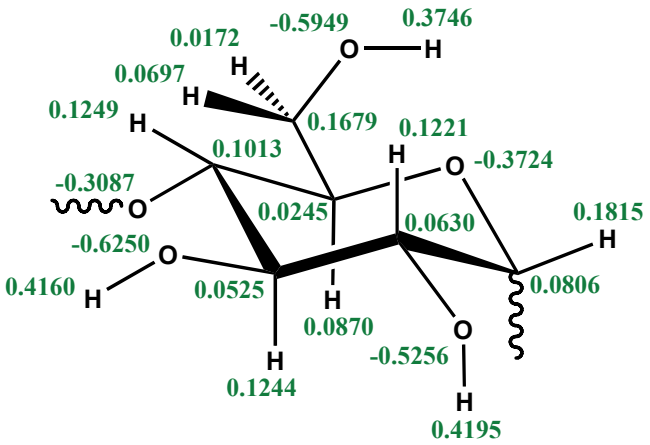


Figure 11.1: Set of RESP Charges for Glucose unit developed by Dr. Iván Beà and employed by Dr. Itziar Maestre and Dr. Miguel de Federico.

11.2.2 Charges by Dr. Javier Pérez Mirón

In principle, the set was calculated using the same conditions employed by Iván Beà and Dr. Itziar Maestre. The only difference is that Dr. Javier Pérez previously did a full conformational search to find the most important conformations. Later, he calculated their weighted contribution to the average set of charges [Fig. 11.2].

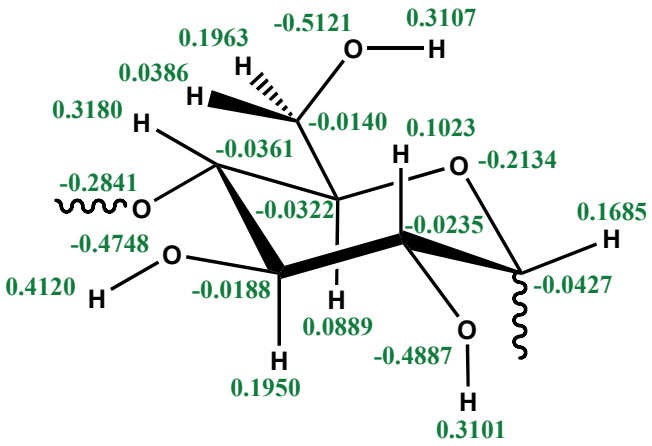


Figure 11.2: Alternative set of RESP Charges for Glucose unit developed by Dr. Javier Pérez.

11.3 Benzene Solvent Model: Units for AMBER 7

At the beginning of this thesis, it was decided to develop our own benzene solvent model. Several reasons made us to move in this direction:

- At the time, there was still no model available for benzene, so we thought that it might be a good idea to create it and offer it to the community of AMBER users.
- Former group member Dr. Xavier Grabuleda successfully developed a solvent model for acetonitrile[1] under the supervision of Peter A. Kollman during his Ph.D. stay at the University of California in San Francisco, so the group already had experience in the area.

[1] Grabuleda, X.; Jaime, C.; Kollman, P.A.; *J. Comput. Chem*. **2000**. *21(10)*. 901-908.

11.3.1 The Set of Charges

We sought for extreme simplicity so our first attempts were focused to define the whole model just by the charges themselves. In this sense, several Quantum Mechanics calculations using Gaussian98 software package were done ranging from medium-sized sets of basis to larger ones also involving different methodologies [Fig. 11.3]:

Names of the sets		*Ab Initio* Methodology		
		HF	MP2	DFT[B3LYP]
Basis	6-31G(d,p)	BNZ4	BNZ7	BNZ1
	6-31++G(d,p)	BNZ5	BNZ8	BNZ2
	6-311++G(d,p)	BNZ6	BNZ9	BNZ3

Figure 11.3: Key of names for the *ab initio* calculations.

The results showed –as was expected- that different conditions affect the final charge values. The tendency points to obtain higher values when the set of basis employed includes more functions. Regarding methodology, the increase in the charges follows the line Density Functional Theory (DFT) B3LYP < Hartree-Fock variational method (HF) <≈ Second Order Møller–Plesset perturbation method (MP2). However, we did not go beyond this point, searching for a better quantum mechanical answers to these values, because it was particularly theoretical and of no use for our direct purposes.

Carbon RESP charges		*Ab Initio* Methodology		
		HF	MP2	DFT[B3LYP]
Basis	6-31G(d,p)	-0.120	-0.120	-0.101
	6-31++G(d,p)	-0.120	-0.121	-0.100
	6-311++G(d,p)	-0.123	-0.123	-0.105

Figure 11.4: Different starting sets of RESP charges for Benzene solvent unit. The molecule is symmetric –D_{6h} point group- and neutral therefore the six carbon atoms have the same charge and opposite to that of the six hydrogens. For this reason only carbon atom charges are included.

Under detailed examination of the table in [Fig. 11.4], two decisions were made:

- The table contains several repeated values although different methodologies have been used to calculate them, thus, the initial set of 9 molecules was reduced to a smaller one: BNZ4, BNZ5 and BNZ7 are equivalent, so they were jointly represented by BNZ4. Similarly, in the cases of BNZ6 and BNZ9, the value of BNZ6 represented both of them.

- It is known that –as Kollman said- the charge calculation by RESP tends to underestimate the PES charges in the 3% to 5% range. Therefore, 3 new groups of 6 sets of charges were created, one of them being 3% and the other one being 5% higher.

	Carbon Charges		
	RESP	RESP + 3%	RESP + 5%
BNZ1	-0.101	-0.104	-0.106
BNZ2	-0.100	-0.103	-0.105
BNZ3	-0.105	-0.108	-0.110
BNZ4	-0.120	-0.124	-0.126
BNZ6	-0.123	-0.127	-0.129
BNZ8	-0.121	-0.125	-0.127

Figure 11.5: Selected sets of RESP charges for Benzene solvent unit including the 3% and 5% perturbed sets.

In order to check the set of charges that best reproduce the macroscopic properties –in principle, density, radial distribution function and isothermal compressibility- Molecular Dynamics calculations were performed: Cubic solvent boxes of 100 angstroms side length involving about 300 molecules were created for the 18 sets of charges [Fig. 11.5]. Long series of equilibrations and samplings under NVT and NPT conditions were performed employing PBC (Periodic Boundary Conditions) and different heating schemas.

The results were not as good as we expected. The calculated density values [Fig. 11.6] were all clearly under the experimental reference[2] value of 0.87865 g/ml. This meaning that the solvent cohesion was poor.

	Charges	Density (g/ml)	
	RESP	average	Std.Dev.
BNZ1	-0.101	0.8237	0.0092
BNZ2	-0.100	0.8299	0.0090
BNZ3	-0.105	0.8320	0.0079
BNZ4	-0.120	0.8373	0.0077
BNZ6	-0.123	0.8460	0.0090
BNZ8	-0.121	0.8441	0.0067

Figure 11.6: Table of Densities obtained from Molecular Dynamics calculations employing the natural set of RESP charges for the Benzene solvent unit.

[2] Several Authors. Edited by Weast, R.C.; *Handbook of Chemistry and Physics*. 56th Edition. CRC Press Inc. **1975-1976**. ISBN 0-87819-455-X.

11.3.2 Conclusions

- The set of charges calculated underestimate the density. All the natural RESP charges, the ones increased by 3% and also the ones increased by 5% are equally insufficient.
- The subtle approach of employing different sets of basis and methodologies in the Quantum Mechanical calculations is absolutely useless.
- The current Benzene model –exclusively based on charges- is excessively simple to model more complex interactions such as π-stacking and van der Waals interactions.

11.3.3 Redefining the Problem

On the grounds of the last three statements, two alternative solutions were proposed:

1) The electrostatic approach: consisting of modelling the π-stacking just by *artificially* increasing the set of charges until they reproduce the experimental data –the worse option-.
2) The van der Waals approach: consisting of modelling the π-stacking by reconsidering again the calculations from different methodologies and introducing a new van der Waals non-bonded interaction parameter in the Carbon-Carbon type –the best option-.

Both solutions were intended to be done just for comparing them and gather material enough for the publication.

11.3.3.1 The Electrostatic Approach

The necessary increment in the charges for achieving the internal cohesion compatible with the experimental density [Fig. 11.7] was estimated by means of linear regression of density vs. RESP charges.

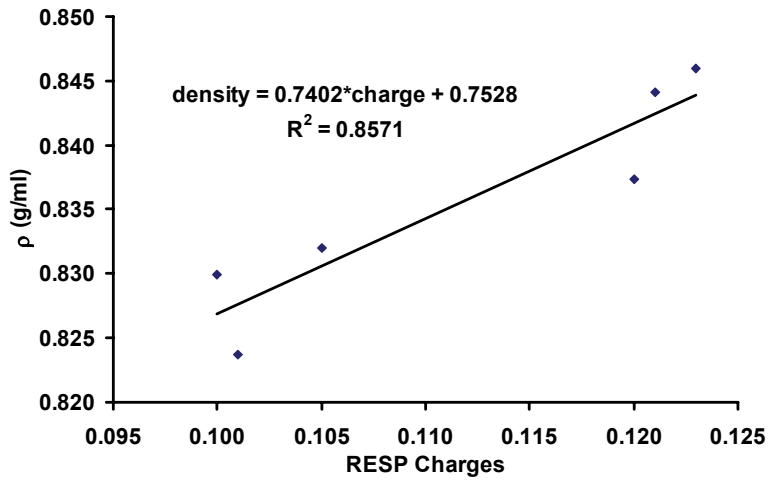


Figure 11.7: Linear regression of density vs. RESP charges.

The value was out of the range but it was just the starting point: the charge value for a density of 0.87865 g/ml is expected to be about 0.169, which according to the table in [Fig. 11.8] is 37% higher than the original RESP charges for the BNZ6 model –that one with the higher natural set of charges-. Anyway, the Molecular Dynamics calculations of the full series were done, finding that the density value for that of RESP + 40% was indeed closer to the one experimental one.

	RESP Charges							
	RESP	**+ 5%**	**+ 10%**	**+ 15%**	**+ 20%**	**+ 30%**	**+ 40%**	**+ 50%**
BNZ6	-0.123	-0.129	-0.135	-0.141	-0.148	-0.160	-0.172	-0.185

Figure 11.8: Table of incremented RESP charges.

The breakdown of charges for the best model is shown hereafter [Fig. 11.9]. As was said, two effects have contributions in the total value: on the one hand, the benzene natural charges. On the other hand, the extra charges added to raise the electrostatic attraction.

	RESP + 40%	-0.172	Total charge	
BNZ6	**RESP**	-0.123	Base charge	[from HF/MP2 6-311++G(d,p)]
	Increment	-0.049	Excess of charge	[accounting for π-stacking]

Figure 11.9: Table of RESP charges analysis for the best set of charges according to the electrostatic approach.

Therefore, the best unit for our tentative solvent model is defined this way [Fig. 11.10]:

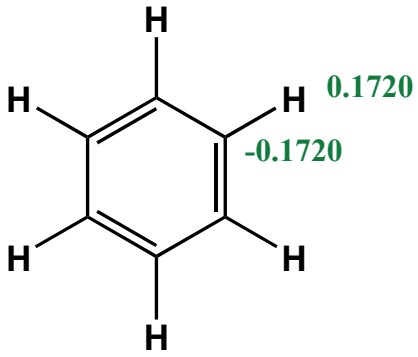


Figure 11.10: Final set of RESP charges for Benzene solvent unit developed in the present work. Only charges for one carbon atom and one hydrogen atom are included.

11.3.3.2 The van der Waals Approach

This part of the study involved the implementation of a van der Waals term in the type C-C to complete the already developed set of charges –the nine different sets under different conditions-.

Unfortunately for us, after more than 6 months of work, when this part of the research was about to start, it was published a paper by Ivo Cacelli, Giorgio Cinacchi, Giacomo Prampolini, and Alessandro Tani in the JACS[3] where they showed their own benzene solvent model. Their study was complete and well developed, so we considered useless going further in this research and abandoned the project in this point, leaving the model in a semi-developed stage.

[3] Cacelli, I.; Cinacchi, G.; Prampolini, G.; Tani A.; *J. Am. Chem. Soc*. **2004**. *126(43)*. 14278-14286.

12 APPENDIX: C-Shell scripts and Java2 Software

12.1 NameGiantCyD.java [Java2]: Improved automatic nomenclature

```
import java.io.*;
import java.util.StringTokenizer;

/**
 *
 * MODO DE USO DEL PROGRAMA.
 *
 * Se lanza desde la línea de comandos invocando la Máquina
 * Virtual de Java (JVM) pasando como argumento el nombre de
 * la clase ya compilada (sin la extensión .class). Los únicos
 * 2 argumentos de la clase deben ser:
 *
 *  1) El nombre del archivo de ángulos de cada conformación, y
 *  2) El tipo de nomenclatura: "itziar" o "kepa".
 *
 * java NameGiantCyD fileName nomenclature > angles.txt
 *
 * El formato de este archivo ha de ser:
 *
 *  1) Tantas líneas como conformaciones.
 *  2) Cada línea tantos ángulos diedros "flip" (O3C4C1O2)
 *           como glucosas tenga la ciclodextrinas.
 *  3) Nada más que no sea lo descrito arriba (ni líneas en
 *           blanco ni otros comentarios)
 *
 * El programa detecta automáticamente el número de estructuras
 * a nombrar y el número de glucosas de la ciclodextrina.
 * La salida la realiza al final de todo el proceso por la STDOUT
 * de modo que se hace necesario redireccionar manualmente el flujo
 * hacia un archivo de texto (angles.txt).
 *
 *
 *                                          Kepa K. Burusco
```

```
 *                                                [30-IX-2007]
 *
 */


public class NameGiantCyD {


      // DECLARACIÓN DE LAS VARIABLES GLOBALES.

      private static int        numCyD, numVar;
      private static double[][] dihedralMatrix;
      private static char[][]   dihedralCharMatrix;
      private static String[]   rawDihedralName;
      private static String[]   vectorOfCyDNames;
      private static int[]      numericCode;
      private static int[][]    conformationCodes;
      private static String     nomenclature;
      private static String     fileName;
      private static String     currLine, workString;
      private static StringTokenizer currLineToken;
      private static BufferedReader  bufferArchive = null;


      // DECLARACIÓN DE LOS MÉTODOS.

      /**
       * @param args
       */

      public static void main(String[] args) throws Exception {

            // Imprime la ayuda por pantalla.
            if ( args[0].trim().equals("-h") || args[0].trim().equals("-help") ) {
                  howTo();
            }
```

```
// Lee los parámetros de entrada para el análisis.
try {
      fileName     = args[0];
      nomenclature = args[1];
} catch(Exception e) {
      System.out.println("\n\tError al intentar cargar los parámetros por");
      System.out.println("\tla línea de comandos.");
      System.gc();
      System.exit(0);
}

/*
 * Carga el archivo con los ángulos O3C4C1O2.
 * La estructura de atrapado maneja la FileNotFound
 * Exception que es obligatorio capturar o delegar
 * cuando se produce.
 *
 * Este paso se hace 2 veces. La primera determina
 * el tamaño de la matriz y la segunda la carga. La
 * estructura de atrapado de errores se omite en el
 * caso 2 ya que si la primera vez no da error, la
 * segunda tampoco lo hará.
 */
try {
      bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
} catch(Exception e) {
      System.out.println("\n\tError al intentar abrir el archivo de los diedros:");
      System.out.println("\tComprobar que la ruta es correcta y el archivo existe.\n");
      System.gc();
      System.exit(0);
}

// Carga las variables número de glucosas y número de conformaciones.
getDimensions(bufferArchive);

try {
      bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
```

```
        } catch(Exception e){}

        // Inicializa dinámicamente la matriz y la carga con los ángulos diedros.
        getData(bufferArchive);

        // Transforma los valores numéricos en la nomenclatura de caracteres.
        anglesIntoCharacters();

        // Compacta la cadena indicando con índices las secuencias de caracteres iguales.
        compactLines();

        /*
         * Como la ciclodextrina es cíclica puede ocurrir que el punto de ruptura de la
         * nomenclatura parta una secuencia del mismo tipo (una secuencia de "s", o de
         * "a", o etc...). Este método comprueba estos casos y los corrige si aparecen.
         * Comprueba la primera y última secuencia de caracteres de la cadena:
         * Si son del mismo tipo suma sus índices, elimina el último campo, actualiza
         * el primero, y rescribe la nueva cadena.
         *
         */

        checkSequenceIntegrity();

        /*
         * Convierte la versión provisional del nombre de la CyD en código numérico y lo
         * carga en un vector, elige el punto de comienzo de la nomenclatura, crea la
         * cadena y la devuelve. Elige el orden correcto y lo carga en el vector solución.
         *
         */

        getBestName();

        /*
         * Imprime por pantalla el vector con los resultados.
         *
         */
```

```
            printVectorOfNames();

            // Llama al recolector de basura, vacía la memoria y termina la ejecución.
            System.gc();
            System.exit(0);

      }// Fin del método main().


      private static void getDimensions(BufferedReader bufferArchive) throws Exception {
            try {
                  numVar = 0;
                  while(true){
                        currLine = bufferArchive.readLine().trim();
                        numVar++;
                        if ( numVar == 1 ){
                              currLineToken = new StringTokenizer(currLine);
                              numCyD = currLineToken.countTokens();
                        }
                  }
            }catch(Exception e){}
            //System.out.println("El número total de conformaciones es: " + numVar);
            //System.out.println("El número de glucosas en la CyD es:   " + numCyD);
            bufferArchive.close();
            bufferArchive = null;
            currLineToken = null;
            currLine      = null;
      }// Fin del método setDimension().


      private static void getData(BufferedReader bufferArchive) throws Exception {
            try {
                  dihedralMatrix = new double[numVar][numCyD];
                  int k = 0;
                  while( k < numVar ){
                        int q = 0;
                        currLine = bufferArchive.readLine().trim();
```

```
                        currLineToken = new StringTokenizer(currLine);
                        while ( q < numCyD ) {
                              workString = currLineToken.nextToken();
                              dihedralMatrix[k][q] = Double.valueOf(workString).doubleValue();
                              //System.out.println("El ángulo " + q + " vale: " + dihedralMatrix[k][q]);
                              q++;
                        }
                        k++;
                  }
            } catch(Exception e) {
                  System.out.println("\n\tError del método getData.\n");
                  System.gc();
                  System.exit(0);
            }
            bufferArchive.close();
            bufferArchive = null;
            currLineToken = null;
            workString    = null;
            currLine      = null;
      }// Fin del método getData().


      private static void anglesIntoCharacters() throws Exception{
            try{

                  dihedralCharMatrix = new char[numVar][numCyD];

                  if ( nomenclature.equals("itziar") ) {
                        int k = 0;
                        while( k < numVar ){
                              int q = 0;
                              while ( q < numCyD ) {
                                    dihedralCharMatrix[k][q] = classifier_IMA(dihedralMatrix[k][q]);
                                    q++;
                              }
                              k++;
                        }
```

```
                } else if ( nomenclature.equals("kepa") ) {
                      int k = 0;
                      while( k < numVar ){
                            int q = 0;
                            while ( q < numCyD ) {
                                  dihedralCharMatrix[k][q] = classifier_KBG(dihedralMatrix[k][q]);
                                  q++;
                            }
                            k++;
                      }
                }

          }catch(Exception e){
                System.out.println("\n\tError del método anglesIntoCharacters.\n");
                System.gc();
                System.exit(0);
          }
    }// Fin del método anglesIntoCharacters().


    private static void compactLines() throws Exception {
          rawDihedralName = new String[numVar];
          char prevChar, currChar, nextChar;
          int k, q, counter;
          workString = "";
          try {
                k = 0;
                while( k < numVar ){
                      q = 0;
                      counter = 1;
                      while ( q < numCyD ) {
                            if ( q <  numCyD - 1 ){
                                  currChar = dihedralCharMatrix[k][q];
                                  nextChar = dihedralCharMatrix[k][q+1];
                                  if ( currChar == nextChar ) {
                                        counter++;
                                  } else {
```

```
                                    workString += ("" + counter + "" + currChar + ",").trim();
                                    counter = 1;
                              }
                        } else if (q == numCyD - 1 ) {
                              currChar = dihedralCharMatrix[k][q];
                              prevChar = dihedralCharMatrix[k][q-1];
                              if ( currChar == prevChar ){
                                    workString += ("" + counter + "" + currChar + "").trim();
                              } else {
                                    workString += ("1" + "" + currChar + "").trim();
                              }
                        }
                        q++;
                  }
                  rawDihedralName[k] = workString;
                  //System.out.println(rawDihedralName[k]);
                  workString = "";
                  k++;
            }
      } catch (Exception e) {
            System.out.println("\n\tError del método compactLines.\n");
            System.gc();
            System.exit(0);
      }
      workString = null;
}// Fin del método compactLines().


private static void checkSequenceIntegrity() throws Exception {

      String firstField, innerFields, lastField;
      int firstIndex, lastIndex, newIndex;
      char firstChar, lastChar;
      int numFields, length, k, q;

      try {
```

```
k = 0;
while ( k < numVar) {

      currLineToken = new StringTokenizer(rawDihedralName[k], ",");
      numFields = currLineToken.countTokens();

      if ( numFields > 2 ) {

            firstField  = currLineToken.nextToken();
            innerFields = "";
            q = 2;
            while ( q < numFields ) {
                  if ( q < numFields - 1 ) {
                        innerFields += ( currLineToken.nextToken() + "," ).trim();
                  } else {
                        innerFields += ( currLineToken.nextToken() + " " ).trim();
                  }
                  q++;
            }
            lastField = currLineToken.nextToken();

            length = firstField.length();
            firstIndex = Integer.valueOf(firstField.substring(0,length -1)).intValue();
            firstChar = firstField.charAt(length -1);
            length = lastField.length();
            lastIndex = Integer.valueOf(lastField.substring(0,length -1)).intValue();
            lastChar = lastField.charAt(length -1);

            if ( firstChar == lastChar ) {
                  newIndex = firstIndex + lastIndex;
                  firstField = ( "" + newIndex + firstChar).trim();
                  rawDihedralName[k] = ( firstField + "," + innerFields ).trim();
            }

      }
      // System.out.println(rawDihedralName[k]);
      k++;
```

```
                }

        } catch (Exception e) {
                System.out.println("\n\tError del método checkSequenceIntegrity.\n");
                System.gc();
                System.exit(0);
        }
        currLineToken = null;
}// Fin del método checkSequenceIntegrity().


private static void getBestName() throws Exception {

        try {

                vectorOfCyDNames = new String[numVar];
                int k = 0;

                while ( k < numVar ){

                        currLineToken = new StringTokenizer(rawDihedralName[k], ",");
                        int numFields = currLineToken.countTokens();
                        numericCode = new int[numFields];
                        conformationCodes = new int[numFields][numFields];

                        // Convierte los tokens en el código de 4 dígitos y carga el vector auxiliar.
                        for ( int q = 0; q < numFields; q++ ) {
                                numericCode[q] = codeTranslator(currLineToken.nextToken());
                        }

                        // Crea la matriz para la ordenación del nombre.
                        int i = 0;
                        while ( i < numFields ) {
                                int j = 0;
                                while ( j < numFields ) {
                                        if (i+j < numFields ) {
                                                conformationCodes[i][j] = numericCode[i+j];
```

```
                                    } else {
                                          conformationCodes[i][j] = numericCode[i+j-numFields];
                                    }
                                    j++;
                              }
                              i++;
                        }

                        /*
                         * Ordena la matriz, elige el nombre de la conformación
                         * y la carga en el vector vectorOfCyDNames.
                         *
                         */

                        sortMatrix(numFields);
                        vectorOfCyDNames[k] = nameConverter(numFields);

                        k++;
                  }

            } catch(Exception e) {
                  System.out.println("\n\tError del método getBestName.\n");
                  System.gc();
                  System.exit(0);
            }
            currLineToken = null;
      }// Fin del método getBestName().


      private static int codeTranslator(String stringCode) throws Exception {

            char confChar;
            char charCode = 'k';
            String index, code;
            int length;
            int numCode = 0;
```

```
try {

      length = stringCode.length();

      if ( length <= 4 ) {

            confChar = stringCode.charAt(length -1);
            index = stringCode.substring(0,length -1).trim();

            if ( nomenclature.equals("itziar") ){

                  if        ( confChar == 's' ) {
                        charCode = '4';
                  } else if ( confChar == 'a' ) {
                        charCode = '3';
                  } else if ( confChar == '+' ) {
                        charCode = '2';
                  } else if ( confChar == '-' ) {
                        charCode = '1';
                  }

            } else if ( nomenclature.equals("kepa") ) {

                  if        ( confChar == 'Z' ) {
                        charCode = '5';
                  } else if ( confChar == 'S' ) {
                        charCode = '4';
                  } else if ( confChar == 'a' ) {
                        charCode = '3';
                  } else if ( confChar == '+' ) {
                        charCode = '2';
                  } else if ( confChar == '-' ) {
                        charCode = '1';
                  }

            }
```

```
                        if ( index.length() == 1 ) {
                              index = ( "00" + index ).trim();
                        } else if ( index.length() == 2) {
                              index = ( " 0" + index ).trim();
                        } else if ( index.length() == 3 ) {
                              index = ( "  " + index ).trim();
                        }

                        code = ( charCode + index + "" ).trim();
                        numCode = Integer.valueOf(code).intValue();

                  } else {
                        System.out.println("\n\tADVERTENCIA del método codeTranslator.\n");
                        System.out.println("\tEste método no admite secuencias de caracteres");
                        System.out.println("\tiguales o superiores a 999 diedros. Revisar el");
                        System.out.println("\tcódigo para admitir estos casos.\n");
                        System.gc();
                        System.exit(0);
                  }

            } catch(Exception e) {
                  System.out.println("\n\tError del método codeTranslator.\n");
                  System.gc();
                  System.exit(0);
            }
            return(numCode);
      }// Fin del método codeTranslator().


      private static void interchangeMatrixVectors(int index1, int index2, int numFields) throws Exception {
            try {
                  int auxBox;
                  for ( int i = 0; i < numFields; i++ ) {
                        auxBox = conformationCodes[index1][i];
                        conformationCodes[index1][i] = conformationCodes[index2][i];
                        conformationCodes[index2][i] = auxBox;
                  }
```

```
            } catch(Exception e) {
                  System.out.println("\n\tError del método interchangeMatrixVectors.\n");
                  System.gc();
                  System.exit(0);
            }
      }// Fin del método interchangeMatrixVectors().


      private static void sortMatrix(int numFields) throws Exception {
            try {
                  int times = numFields;
                  int tmpTimes = 0;
                  int maxValue;

                  for ( int i = 0; i < numFields; i++ ){
                        // Ordenación.
                        for ( int j = 0; j < times; j++ ) {
                              for ( int k = j + 1; k < times; k++ ) {
                                    if ( conformationCodes[k][i] > conformationCodes[j][i] ) {
                                          interchangeMatrixVectors(j, k, numFields);
                                    }
                              }
                        }
                        // Busco el valor más alto.
                        maxValue = conformationCodes[0][i];
                        for (int q = 1; q < times; q++){
                              if ( conformationCodes[q][i] > maxValue ) {
                                    maxValue = conformationCodes[q][i];
                              }
                        }
                        // Veo cuánto se repite.
                        tmpTimes = 0;
                        for ( int n = 0; n < times; n++){
                              if ( conformationCodes[n][i] == maxValue ) {
                                    tmpTimes++;
                              }
                        }
```

```
                    times = tmpTimes;
                    if ( times == 1 ) { break; }

              }

        } catch(Exception e) {
              System.out.println("\n\tError del método sortMatrix.\n");
              System.gc();
              System.exit(0);
        }
  }// Fin del método sortMatrix().


  private static String nameConverter(int numFields) throws Exception {
        currLine = "";
        int index, numCode;
        char charCode = 'k';

        try {

              for (int i = 0; i < numFields; i++) {
                    workString = ("" + conformationCodes[0][i]).trim();
                    numCode = Integer.valueOf("" + workString.charAt(0)).intValue();
                    index   = Integer.valueOf("" + workString.substring(1)).intValue();

                    if ( nomenclature.equals("itziar") ) {

                          if        ( numCode == 4 ) {
                                charCode = 's';
                          } else if ( numCode == 3 ) {
                                charCode = 'a';
                          } else if ( numCode == 2 ) {
                                charCode = '+';
                          } else if ( numCode == 1 ) {
                                charCode = '-';
                          }
```

```
                        } else if ( nomenclature.equals("kepa") ) {

                              if        ( numCode == 5 ) {
                                    charCode = 'Z';
                              } else if ( numCode == 4 ) {
                                    charCode = 'S';
                              } else if ( numCode == 3 ) {
                                    charCode = 'a';
                              } else if ( numCode == 2 ) {
                                    charCode = '+';
                              } else if ( numCode == 1 ) {
                                    charCode = '-';
                              }

                        }


                        if ( index == 1) {
                              workString = ( " " + charCode ).trim();
                        } else {
                              workString = ( " " + index + charCode ).trim();
                        }

                        if ( i < numFields -1 ){
                              currLine += ( " " + workString + ",").trim();
                        } else {
                              currLine += ( " " + workString + " ").trim();
                        }
                  }

            } catch(Exception e) {
                  System.out.println("\n\tError del método nameConverter.\n");
                  System.gc();
                  System.exit(0);
            }
            workString = null;
            return(currLine);
```

```
	}// Fin del metodo nameConverter().


	private static char classifier_IMA(double angle) throws Exception {
		char charValue = ' ';
		try {
			if	( angle >  -135.00000 && angle <  -60.00000 ){
				charValue = '-';
			} else if ( angle >=  -60.00000 && angle <=  60.00000 ){
				charValue = 's';
			} else if ( angle >    60.00000 && angle <  135.00000 ){
				charValue = '+';
			} else {
				charValue = 'a';
			}
		} catch(Exception e) {
			System.out.println("\n\tError del método classifier.\n");
			System.gc();
			System.exit(0);
		}
		return(charValue);
	}// Fin del método anglesIntoCharacters_IMA().


	private static char classifier_KBG(double angle) throws Exception {
		char charValue = ' ';
		try {
			if	( angle > -155.00000 && angle < -60.00000 ){
				charValue = '-';
			} else if ( angle >= -60.00000 && angle <   0.00000 ){
				charValue = 'Z';
			} else if ( angle >=   0.00000 && angle <  80.00000 ){
				charValue = 'S';
			} else if ( angle >=  80.00000 && angle < 125.00000 ){
				charValue = '+';
			} else {
				charValue = 'a';
```

```
                }
        } catch(Exception e) {
                System.out.println("\n\tError del método classifier.\n");
                System.gc();
                System.exit(0);
        }
        return(charValue);
}// Fin del método anglesIntoCharacters_KBG().


private static void printVectorOfNames() throws Exception {
        try{
                for (int i = 0; i < numVar; i++) {
                      System.out.println(vectorOfCyDNames[i]);
                }
        }catch(Exception e) {
                System.out.println("\n\tError del método printVectorOfNames.\n");
                System.gc();
                System.exit(0);
        }
}// Fin del método printVectorOfNames().


private static void howTo() {

        System.out.println(

        "\n\tMODO DE USO DEL PROGRAMA.\n\n" +

        "Se lanza desde la línea de comandos invocando la Máquina\n" +
        "Virtual de Java (JVM) pasando como argumento el nombre de\n" +
        "la clase ya compilada (sin la extensión .class). Los únicos\n" +
        "2 argumento de la clase deben ser:\n" +
        " 1) El nombre del archivo de ángulos de cada conformación, y\n" +
        " 2) El tipo de nomeclatura: \"itziar\" ó \"kepa\".\n\n" +

        "java NameGiantCyD fileName nomenclature > angles.txt\n\n" +
```

```
            "El formato de este archivo ha de ser:\n" +
            " 1) Tantas líneas como conformaciones.\n" +
            " 2) Cada línea tantos ángulos diedros \"flip\" (O3C4C1O2)\n" +
            "          como glucosas tenga la ciclodextrinas.\n" +
            " 3) Nada más que no sea lo descrito arriba (ni líneas en\n" +
            "          blanco ni otros comentarios)\n\n" +

            "El programa detecta automáticamente el número de estructuras\n" +
            "a nombrar y el número de glucosas de la ciclodextrina.\n" +
            "La salida la realiza al final de todo el proceso por la STDOUT\n" +
            "de modo que se hace necesario redireccionar manualmente el flujo\n" +
            "hacia un archivo de texto (angles.txt).\n\n" +

            "\t\t\t\t\tKepa K. Burusco\n" +
            "\t\t\t\t\t[5-X-2007]\n\n");

            System.gc();
            System.exit(0);


      }// Fin del método howTo().


}// Fin de la Clase NameGiantCyD
```

12.2 SortDihedral.java [Java2]: Removes CDs cyclic symmetry before PCA

```
import java.io.*;
import java.util.StringTokenizer;

/**
```

```
 *
 * MODO DE USO DEL PROGRAMA.
 *
 * Se lanza desde la línea de comandos invocando la Máquina
 * Virtual de Java (JVM) pasando como argumento el nombre de
 * la clase ya compilada (sin la extensión .class). Los únicos
 * 2 argumentos de la clase deben ser:
 *
 *  1) El nombre del archivo de ángulos de cada conformación, y
 *  2) El tipo de ordenación a seguir: "itziar" ó "kepa".
 *
 * java SortDihedral fileName ordenacion > angles.txt
 *
 * El formato de este archivo ha de ser:
 *
 *  1) Tantas líneas como conformaciones.
 *  2) Cada línea tantos ángulos diedros "flip" (O3C4C1O2)
 *           como glucosas tenga la ciclodextrinas.
 *  3) Nada más que no sea lo descrito arriba (ni líneas en
 *           blanco ni otros comentarios)
 *
 * El programa detecta automáticamente el número de estructuras
 * a nombrar y el número de glucosas de la ciclodextrina.
 * La salida la realiza al final de todo el proceso por la STDOUT
 * de modo que se hace necesario redireccionar manualmente el flujo
 * hacia un archivo de texto (angles.txt).
 *
 *
 *                                              Kepa K. Burusco
 *                                              [7-XII-2007]
 *
 */


public class SortDihedral {

      // DECLARACIÓN DE LAS VARIABLES GLOBALES.
```

```
private static int        numCyD, numVar;
private static double[][] dihedralMatrix, doubleMatrix;
private static int[][]    dihedralIntMatrix, intMatrix;
private static String     nomenclature, fileName, currLine, workString;
private static StringTokenizer currLineToken;
private static BufferedReader  bufferArchive = null;


// DECLARACIÓN DE LOS MÉTODOS.

/**
 * @param args
 */

public static void main(String[] args) throws Exception {

      // Imprime la ayuda por pantalla.
      if ( args[0].trim().equals("-h") || args[0].trim().equals("-help") ) {
            howTo();
      }

      // Lee los parámetros de entrada para el análisis.
      try {
            fileName     = args[0];
            nomenclature = args[1];
      } catch(Exception e) {
            System.out.println("\n\tError al intentar cargar los parámetros por");
            System.out.println("\tla línea de comandos.");
            System.gc();
            System.exit(0);
      }

      /*
       * Carga el archivo con los ángulos O3C4C1O2.
       * La estructura de atrapado maneja la FileNotFound
       * Exception que es obligatorio capturar o delegar
```

```
         * cuando se produce.
         *
         * Este paso se hace 2 veces. La primera determina
         * el tamaño de la matriz y la segunda la carga. La
         * estructura de atrapado de errores se omite en el
         * caso 2 ya que si la primera vez no da error, la
         * segunda tampoco lo hará.
         */
        try {
              bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
        } catch(Exception e) {
              System.out.println("\n\tError al intentar abrir el archivo de los diedros:");
              System.out.println("\tComprobar que la ruta es correcta y el archivo existe.\n");
              System.gc();
              System.exit(0);
        }

        // Carga las variables número de glucosas y número de conformaciones.
        getDimensions(bufferArchive);

        try {
              bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
        } catch(Exception e){}

        // Inicializa dinámicamente la matriz y la carga con los ángulos diedros.
        getData(bufferArchive);

        // Transforma los valores numéricos en la nomenclatura de caracteres.
        anglesIntoCode();

        // Ordena las matrices comenzando por la secuencia de mayor prioridad.
        sortDihedralMatrix();

        // Crea las líneas (string) ordenadas, carga el vector, y lo imprime por pantalla.
        printSortedDihedral();

        // Llama al recolector de basura, vacía la memoria y termina la ejecución.
```

```
            System.gc();
            System.exit(0);

      }// Fin del método main().


      private static void getDimensions(BufferedReader bufferArchive) throws Exception {
            try {
                  numVar = 0;
                  while(true){
                        currLine = bufferArchive.readLine().trim();
                        numVar++;
                        if ( numVar == 1 ){
                              currLineToken = new StringTokenizer(currLine);
                              numCyD = currLineToken.countTokens();
                        }
                  }
            }catch(Exception e){}
            //System.out.println("El número total de conformaciones es: " + numVar);
            //System.out.println("El número de glucosas en la CyD es:   " + numCyD);
            bufferArchive.close();
            bufferArchive = null;
            currLineToken = null;
            currLine      = null;
      }// Fin del método setDimension().


      private static void getData(BufferedReader bufferArchive) throws Exception {
            try {
                  dihedralMatrix = new double[numVar][numCyD];
                  int k = 0;
                  while( k < numVar ){
                        int q = 0;
                        currLine = bufferArchive.readLine().trim();
                        currLineToken = new StringTokenizer(currLine);
                        while ( q < numCyD ) {
                              workString = currLineToken.nextToken();
```

```
                        dihedralMatrix[k][q] = Double.valueOf(workString).doubleValue();
                        //System.out.println("El ángulo " + q + " vale: " + dihedralMatrix[k][q]);
                        q++;
                  }
                  k++;
            }
      } catch(Exception e) {
            System.out.println("\n\tError del método getData.\n");
            System.gc();
            System.exit(0);
      }
      bufferArchive.close();
      bufferArchive = null;
      currLineToken = null;
      workString    = null;
      currLine      = null;
}// Fin del método getData.


private static void anglesIntoCode() throws Exception{
      try{

            dihedralIntMatrix = new int[numVar][numCyD];

            if ( nomenclature.equals("itziar") ) {
                  int k = 0;
                  while( k < numVar ){
                        int q = 0;
                        while ( q < numCyD ) {
                              dihedralIntMatrix[k][q] = classifier_IMA(dihedralMatrix[k][q]);
                              q++;
                        }
                        k++;
                  }
            } else if ( nomenclature.equals("kepa") ) {
                  int k = 0;
                  while( k < numVar ){
```

```
                              int q = 0;
                              while ( q < numCyD ) {
                                    dihedralIntMatrix[k][q] = classifier_KBG(dihedralMatrix[k][q]);
                                    q++;
                              }
                              k++;
                        }
                  }

            }catch(Exception e){
                  System.out.println("\n\tError del método anglesIntoCharacters.\n");
                  System.gc();
                  System.exit(0);
            }
      }// Fin del método anglesIntoCode().


      private static void sortDihedralMatrix() throws Exception {
            try {
                  int[]    auxInt    = new int[numCyD];
                  double[] auxDouble = new double[numCyD];
                  int n = 0;
                  while( n < numVar ){
                        int q = 0;
                        while( q < numCyD ){
                              auxInt[q]    = dihedralIntMatrix[n][q];
                              auxDouble[q] = dihedralMatrix[n][q];
                              q++;
                        }
                        auxDouble = sortMatrix(auxInt, auxDouble);
                        int m = 0;
                        while( m < numCyD ){
                              dihedralMatrix[n][m] = auxDouble[m];
                              m++;
                        }
                        n++;
                  }
```

```
            }catch(Exception e){
                  System.out.println("\n\tError del método sortDihedralMatrix.\n");
                  System.gc();
                  System.exit(0);
            }
      }// Fin del método sortDihedralMatrix(),


      private static void printSortedDihedral() throws Exception {
            try{
                  for (int i = 0; i < numVar; i++ ){
                        currLine = "";
                        for (int j = 0; j < numCyD; j++ ){
                              currLine += stringFormat(dihedralMatrix[i][j]);
                        }
                        System.out.println(currLine);
                  }
            }catch(Exception e){
                  System.out.println("\n\tError del método printSortedDihedral.\n");
                  System.gc();
                  System.exit(0);
            }
      }// Fin del método printSortedDihedral().


      private static double[] sortMatrix(int[] auxInt, double[] auxDouble) throws Exception {
            intMatrix    = new int[numCyD][numCyD];
            doubleMatrix = new double[numCyD][numCyD];

            // Cargo las matrices de enteros y doubles.
            try{
                  int i = 0;
                  while ( i < numCyD ) {
                        int j = 0;
                        while ( j < numCyD ) {
                              if ( i+j < numCyD ) {
                                    intMatrix[i][j]    = auxInt[i+j];
```

```
                        doubleMatrix[i][j] = auxDouble[i+j];
                    } else {
                        intMatrix[i][j]    = auxInt[i+j-numCyD];
                        doubleMatrix[i][j] = auxDouble[i+j-numCyD];
                    }
                    j++;
                }
                i++;
            }
        }catch(Exception e){
            System.out.println("\n\tError 1 del método sortMatrix.");
            System.out.println("\tFallo al cargar las matrices auxiliares.\n");
            System.gc();
            System.exit(0);
        }

        // Ordeno las matrices de enteros y doubles.
        try{
            int times = numCyD;
            int tmpTimes = 0;
            int maxValue;

            for ( int i = 0; i < numCyD; i++ ){
                // Ordenación.
                for ( int j = 0; j < times; j++ ) {
                    for ( int k = j + 1; k < times; k++ ) {
                        if ( intMatrix[k][i] > intMatrix[j][i] ) {
                            interchangeIntMatrixVectors(j, k);
                            interchangeDoubleMatrixVectors(j, k);
                        }
                    }
                }
                // Busco el valor más alto.
                maxValue = intMatrix[0][i];
                for (int q = 1; q < times; q++){
                    if ( intMatrix[q][i] > maxValue ) {
                        maxValue = intMatrix[q][i];
```

```
                        }
                    }
                    // Veo cuánto se repite.
                    tmpTimes = 0;
                    for ( int n = 0; n < times; n++){
                        if ( intMatrix[n][i] == maxValue ) {
                            tmpTimes++;
                        }
                    }
                    times = tmpTimes;
                    if ( times == 1 ) { break;}
                }
            } catch(Exception e) {
                System.out.println("\n\tError 2 del método sortMatrix.");
                System.out.println("\tFallo al ordenar las matrices auxiliares.\n");
                System.gc();
                System.exit(0);
            }

            // Cargo el vector solución con la primera fila de la matriz auxDouble.
            try{
                for (int c = 0; c < numCyD; c++) {
                    auxDouble[c] = doubleMatrix[0][c];
                }
            }catch(Exception e){
                System.out.println("\n\tError 3 del método sortMatrix.");
                System.out.println("\tFallo al cargar el vector solución.\n");
                System.gc();
                System.exit(0);
            }
            return(auxDouble);
        }// Fin del método sortMatrix().
```

```
	private static void interchangeIntMatrixVectors(int index1, int index2) throws Exception {
		try {
			int auxBox;
			for ( int i = 0; i < numCyD; i++ ) {
				auxBox = intMatrix[index1][i];
				intMatrix[index1][i] = intMatrix[index2][i];
				intMatrix[index2][i] = auxBox;
			}
		} catch(Exception e) {
			System.out.println("\n\tError del método int interchangeIntMatrixVectors.\n");
			System.gc();
			System.exit(0);
		}
	}// Fin del método int interchangeIntMatrixVectors().


	private static void interchangeDoubleMatrixVectors(int index1, int index2) throws Exception {
		try {
			double auxBox;
			for ( int i = 0; i < numCyD; i++ ) {
				auxBox = doubleMatrix[index1][i];
				doubleMatrix[index1][i] = doubleMatrix[index2][i];
				doubleMatrix[index2][i] = auxBox;
			}
		} catch(Exception e) {
			System.out.println("\n\tError del método double interchangeDoubleMatrixVectors.\n");
			System.gc();
			System.exit(0);
		}
	}// Fin del método double interchangeDoubleMatrixVectors().


	private static int classifier_IMA(double angle) throws Exception {
		int numValue = 0;
		try {
			if	( angle >  -135.00000 && angle <  -60.00000 ){
				numValue = 1;
```

```
                } else if ( angle >=  -60.00000 && angle <=  60.00000 ){
                      numValue = 4;
                } else if ( angle >    60.00000 && angle <  135.00000 ){
                      numValue = 2;
                } else {
                      numValue = 3;
                }
          } catch(Exception e) {
                System.out.println("\n\tError del método classifier_IMA.\n");
                System.gc();
                System.exit(0);
          }
          return(numValue);
    }// Fin del método classifier_IMA().


    private static int classifier_KBG(double angle) throws Exception {
          int numValue = 0;
          try {
                if        ( angle > -155.00000 && angle < -60.00000 ){
                      numValue = 1;
                } else if ( angle >= -60.00000 && angle <   0.00000 ){
                      numValue = 5;
                } else if ( angle >=   0.00000 && angle <  80.00000 ){
                      numValue = 4;
                } else if ( angle >=  80.00000 && angle < 125.00000 ){
                      numValue = 2;
                } else {
                      numValue = 3;
                }
          } catch(Exception e) {
                System.out.println("\n\tError del método classifier_KBG.\n");
                System.gc();
                System.exit(0);
          }
          return(numValue);
    }// Fin del método classifier_KBG().
```

```
private static String stringFormat(double dihedralValue) {
      int n = 4;
      int dotPlace = 0;
      String formatedString = "          " + dihedralValue + "0000000000";
      try {
            dotPlace = formatedString.lastIndexOf(".");
            int startPoint  = dotPlace - 7;
            int endPoint    = dotPlace + n + 1;
            formatedString  = formatedString.substring(startPoint, endPoint);
      } catch (Exception e) {
            System.out.println("\n\tError del método stringFormat.\n");
            System.gc();
            System.exit(0);
      }
      return(formatedString);
}// Fin del método formatedString().


private static void howTo() {

      System.out.println(

      "\n\tMODO DE USO DEL PROGRAMA.\n\n" +

      "Se lanza desde la línea de comandos invocando la Máquina\n" +
      "Virtual de Java (JVM) pasando como argumento el nombre de\n" +
      "la clase ya compilada (sin la extensión .class). Los únicos\n" +
      "2 argumento de la clase deben ser:\n" +
      " 1) El nombre del archivo de ángulos de cada conformación, y\n" +
      " 2) El tipo de ordenacion a seguir: \"itziar\" ó \"kepa\".\n\n" +

      "java SortDihedral fileName ordenacion > angles.txt\n\n" +

      "El formato de este archivo ha de ser:\n" +
      " 1) Tantas líneas como conformaciones.\n" +
```

```
            " 2) Cada línea tantos ángulos diedros \"flip\" (O3C4C1O2)\n" +
            "          como glucosas tenga la ciclodextrinas.\n" +
            " 3) Nada más que no sea lo descrito arriba (ni líneas en\n" +
            "          blanco ni otros comentarios)\n\n" +

            "El programa detecta automáticamente el número de estructuras\n" +
            "a nombrar y el número de glucosas de la ciclodextrina.\n" +
            "La salida la realiza al final de todo el proceso por la STDOUT\n" +
            "de modo que se hace necesario redireccionar manualmente el flujo\n" +
            "hacia un archivo de texto (angles.txt).\n\n" +

            "\t\t\t\t\tKepa K. Burusco\n" +
            "\t\t\t\t\t[7-XII-2007]\n\n");

            System.gc();
            System.exit(0);

      }// Fin del método howTo().


}// Fin de la clase SortDihedral().
```

12.3 MarkovChainAnalysis.java [Java2]: Nomenclature Statistics and Markov Transition Matrix

```
import java.io.*;
import java.text.DecimalFormat;

/**
 *
 * MODO DE USO DEL PROGRAMA.
 *
 * Se lanza desde la línea de comandos invocando la Máquina
 * Virtual de Java (JVM) pasando como argumento el nombre de
 * la clase ya compilada (sin la extensión .class). El único
```

```
 * argumento de la clase debe ser el nombre del archivo con
 * la cadena de nombres:
 *
 * java MarkovChainAnalysis fileNameSequence > MarkovResults.txt
 *
 * El formato de este archivo ha de ser:
 *
 *  1) Una conformación por línea.
 *  2) Una sola columna.
 *
 * El programa detecta automáticamente el tamaño de la columna.
 * La salida la realiza al final de todo el proceso por la STDOUT
 * de modo que se hace necesario redireccionar manualmente el flujo
 * hacia un archivo de texto (MarkovResults.txt).
 *
 *
 *                                              Kepa K. Burusco
 *                                              [20-II-2009]
 *
 */


public class MarkovChainAnalysis {


      // DECLARACIÓN DE LAS VARIABLES GLOBALES.

      private static int        numSteps, numConf;
      private static String[]   tempArray;
      private static String[]   vectorOfSteps;
      private static String[]   vectorOfConformations;
      private static int[]      vectorOfDensityInt;
      private static double[]   vectorOfDensityDouble;
      private static double[]   equilibriumVectorDouble;
      private static int[][]    markovTransMatrixInt;
      private static double[][] markovTransMatrixDouble;
      private static int        currIndex, nextIndex;
```

```
	private static double     theta, correlation;
	private static String     fileName;
	private static BufferedReader  bufferArchive = null;



	// DECLARACIÓN DE LOS MÉTODOS.

	/**
	 * @param args
	 */


	public static void main(String[] args) throws Exception {

		//    Imprime la ayuda por pantalla.
		if ( args[0].trim().equals("-h") || args[0].trim().equals("-help") ) {
			howTo();
		}

		//    Lee los parámetros de entrada para el análisis.
		try {
			fileName     = args[0];
		} catch(Exception e) {
			System.out.println("\n\tError al intentar cargar los parámetros por");
			System.out.println("\tla línea de comandos.");
			System.gc();
			System.exit(0);
		}

		/*
		 * Carga el archivo con con las conformaciones.
		 * La estructura de atrapado maneja la FileNotFound
		 * Exception que es obligatorio capturar o delegar
		 * cuando se produce.
		 *
		 * Este paso se hace 2 veces. La primera determina
```

```
         * el tamaño de la columna y la segunda la carga. La
         * estructura de atrapado de errores se omite en el
         * caso 2 ya que si la primera vez no dió error, la
         * segunda tampoco lo hará.
         */
        try {
              bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
        } catch(Exception e) {
              System.out.println("\n\tError al intentar abrir el archivo de conformaciones:");
              System.out.println("\tComprobar que la ruta es correcta y el archivo existe.\n");
              System.gc();
              System.exit(0);
        }

        //    Carga las variables número de glucosas y número de conformaciones.
        getDimension(bufferArchive);

        try {
              bufferArchive = new BufferedReader(new FileReader(new File(fileName)));
        } catch(Exception e){}

        // Inicializa dinámicamente el vector y lo carga con las conformaciones (steps).
        getData(bufferArchive);

        // Crea el vector de conformaciones.
        manageConformations();

        // Crea la Matriz de Transiciones de Markov.
        markovMatrix();

        // Calcula los valores porcentuales y estadísticos.
        statistics();

        // Imprime resultados por la salida estandar (STDOUT).
        printResults();

        // Llama al recolector de basura, vacía la memoria y termina la ejecución.
```

```
            System.gc();
            System.exit(0);

      }// Fin del método main().


      private static void getDimension(BufferedReader bufferArchive) throws Exception {
            try {
                  numSteps = 0;
                  while(true){
                        bufferArchive.readLine().trim();
                        numSteps++;
                  }
            }catch(Exception e){}
            bufferArchive.close();
            bufferArchive = null;
      }// Fin del método getDimension().

      private static void getData(BufferedReader bufferArchive) throws Exception {
            try {
                  vectorOfSteps = new String[numSteps];
                  int k = 0;
                  while( k < numSteps ){
                        vectorOfSteps[k] = bufferArchive.readLine().trim();
                        //System.out.println("El confórmero del ciclo " + k + " es: " + vectorOfSteps[k]);
                        k++;
                  }
            } catch(Exception e) {
                  System.out.println("\n\tError del método getData.\n");
                  System.gc();
                  System.exit(0);
            }
            bufferArchive.close();
            bufferArchive = null;
      }// Fin del método getData().

      private static void manageConformations() throws Exception {
```

```
try {
      tempArray = new String[numSteps];
      for (int n = 0 ; n < numSteps ; n++) {
            tempArray[n] = vectorOfSteps[n];
      }
      // Ordenación del vector por el método de la burbuja.
      for (int j = 0 ; j < numSteps; j++){
            for (int k = j + 1 ; k < numSteps ; k++) {
                  if (tempArray[k].compareTo(tempArray[j]) < 0) {
                        String tmpStr = tempArray[j];
                        tempArray[j]  = tempArray[k];
                        tempArray[k]  = tmpStr;
                  }
            }
      }
} catch (Exception e) {
      System.out.println("\n\tError del método manageConformations. Parte I.\n");
      System.gc();
      System.exit(0);
}// Fin Parte I del método.

try {
      // Determina cuántos confórmeros diferentes hay.
      numConf = 1;
      int k = 0;
      for (int j = 0 ; j < numSteps -1 ; j++){
            k = j + 1;
            if ( !tempArray[k].equals(tempArray[j]) ) {
                  numConf++;
            }
      }
} catch (Exception e) {
      System.out.println("\n\tError del método manageConformations. Parte II.\n");
      System.gc();
      System.exit(0);
}// Fin Parte II del método.
```

```
try {
      // Crea los vectores de conformaciones y densidades.
      vectorOfConformations = new String[numConf];
      vectorOfDensityInt        = new int[numConf];
      int i = 0;
      int j = 0;
      int confNumber = 1;
      vectorOfConformations[confNumber -1] = tempArray[0];
      vectorOfDensityInt[confNumber -1] = 1;
      while (i < numSteps - 1){
            j = i + 1;
            if ( tempArray[j].equals(tempArray[i]) ) {
                  vectorOfDensityInt[confNumber -1]++;
            } else {
                  confNumber++;
                  vectorOfConformations[confNumber -1] = tempArray[j];
                  vectorOfDensityInt[confNumber -1] = 1;
            }
            i++;
      }
} catch (Exception e) {
      System.out.println("\n\tError del método manageConformations. Parte III.\n");
      System.gc();
      System.exit(0);
}// Fin Parte III del método.

try {
      // Ordena los vectores de conformaciones y densidades por el método de la burbuja.
      for (int j = 0 ; j < numConf; j++){
            for (int k = j + 1 ; k < numConf ; k++) {
                  if ( vectorOfDensityInt[k] > vectorOfDensityInt[j] ) {
                        int q = vectorOfDensityInt[j];
                        vectorOfDensityInt[j] = vectorOfDensityInt[k];
                        vectorOfDensityInt[k] = q;
                        String tmpStr = vectorOfConformations[j];
                        vectorOfConformations[j] = vectorOfConformations[k];
                        vectorOfConformations[k] = tmpStr;
```

```
                              }
                        }
                  }
            }catch (Exception e){
                  System.out.println("\n\tError del método manageConformations. Parte IV.\n");
                  System.gc();
                  System.exit(0);
            }// Fin Parte IV del método.

      }// Fin del método manageConformations().

      private static void markovMatrix() throws Exception {
            try {
                  // Inicializa la matriz de Markov con ceros.
                  markovTransMatrixInt = new int[numConf][numConf];
                  for (int i = 0 ; i < numConf; i++){
                        for (int j = 0 ; j < numConf ; j++) {
                              markovTransMatrixInt[i][j]= 0;
                        }
                  }
                  // Carga la matriz de Markov a partir de la secuencia.
                  int q = 0;
                  while ( q < numSteps -1 ){
                        currIndex = getIndex(vectorOfSteps[q]);
                        nextIndex = getIndex(vectorOfSteps[q+1]);
                        markovTransMatrixInt[currIndex][nextIndex]++;
                        q++;
                  }
            } catch (Exception e) {
                  System.out.println("\n\tError del método markovMatrix.\n");
                  System.gc();
                  System.exit(0);
            }
      }// Fin del método markovMatrix().

      private static void statistics() throws Exception {
            try {
```

```
                    // Vector de densidades (Reales en tanto por uno).
                    vectorOfDensityDouble = new double[numConf];
                    for (int i = 0; i < numConf; i++) {
                          vectorOfDensityDouble[i] = Double.parseDouble(" " +
vectorOfDensityInt[i])/Double.parseDouble(" " + numSteps);
                    }
                    // Matriz de transición de Markov (Reales en tanto por uno).
                    markovTransMatrixDouble = new double[numConf][numConf];
                    for (int i = 0 ; i < numConf; i++) {
                          int normal = 0;
                          for (int j = 0 ; j < numConf ; j++) {
                                normal += markovTransMatrixInt[i][j];
                          }
                          for (int j = 0 ; j < numConf ; j++) {
                                markovTransMatrixDouble[i][j] = Double.parseDouble(" " +
markovTransMatrixInt[i][j])/Double.parseDouble(" " + normal);
                          }
                    }
             } catch (Exception e) {
                          System.out.println("\n\tError del método statistics. Parte I.\n");
                          System.gc();
                          System.exit(0);
             }// Fin de la Parte I.

             try {
                    /*
                     * Calcula el vector de densidad a partir de la Matriz de Markov
                     * obtenida a partir de la cadena, y del vector de densidad obtenido
                     * del análisis estadístico. Si el vector que sale de este producto
                     * es igual -o comparable- al vector densidad obtenido del análisis
                     * estadístico entonces el sistema está en equilibrio.
                     */
                    equilibriumVectorDouble = new double[numConf];
                    for (int i = 0 ; i < numConf; i++) {
                          equilibriumVectorDouble[i] = 0.0;
                          for (int j = 0 ; j < numConf ; j++) {
                                equilibriumVectorDouble[i] += markovTransMatrixDouble[j][i]*vectorOfDensityDouble[j];
```

```
                }
            }

        } catch (Exception e) {
            System.out.println("\n\tError del método statistics. Parte II.\n");
            System.gc();
            System.exit(0);
        }// Fin de la Parte II.

        try {
            /*
             * Calcula el producto escalar (correlación) y el ángulo entre los
             * vectores de densidad -el obtenido por análisis estadístico y el
             * que sale del producto de la matriz de transiciones por el obtenido
             * por análisis estadístico-.
             *
             */
            double average_St = 0.0;
            double average_Eq = 0.0;
            for (int i = 0; i < numConf; i++ ) {
                average_St += vectorOfDensityDouble[i]/numConf;
                average_Eq += equilibriumVectorDouble[i]/numConf;
            }

            double[] array_St = new double[numConf];
            double[] array_Eq = new double[numConf];
            for (int i = 0; i < numConf; i++ ) {
                array_St[i] = vectorOfDensityDouble[i] - average_St;
                array_Eq[i] = equilibriumVectorDouble[i] - average_Eq;
            }

            double prod_StEq = 0.0;
            double mod_St = 0.0;
            double mod_Eq = 0.0;
            for (int i = 0; i < numConf; i++ ) {
                prod_StEq += array_St[i]*array_Eq[i];
                mod_St    += array_St[i]*array_St[i];
```

```
                        mod_Eq    += array_Eq[i]*array_Eq[i];
                  }
                  mod_St = Math.sqrt(mod_St);
                  mod_Eq = Math.sqrt(mod_Eq);

                  correlation = prod_StEq/(mod_St*mod_Eq);
                  theta = Math.acos(correlation);
                  if (theta <= 0.000001 && theta >= -0.000001) {
                        theta  = 0.000000;
                  }

            } catch (Exception e) {
                  System.out.println("\n\tError del método statistics. Parte III\n");
                  System.gc();
                  System.exit(0);
            }// Fin de la Parte III.

      }// Fin del método statistics().

      private static int getIndex(String nameConf) throws Exception {
            int index = 0;
            try {
                  /*
                   * Este método determina la posición de la conformación
                   * en el vector de conformaciones y por tanto en la matriz
                   * de transiciones. Con este método determino el par
                   * cartesiano (x,y) para ir llenando la matriz de
                   * transiciones.
                   *
                   */
                  while (index < numConf && !vectorOfConformations[index].equals(nameConf) ){
                        index++;
                  }
            }catch (Exception e){
                  System.out.println("\n\tError del método getIndex.\n");
                  System.gc();
                  System.exit(0);
```

```
        }
        return(index);
    }// Fin del metodo getIndex().

    private static void printResults() throws Exception {
        try{

            String string = new String();
            System.out.println("\n");
            System.out.println("   ------------------------------------------------");
            System.out.println("         Markov Analysis and General Fitness      ");
            System.out.println("   ------------------------------------------------");
            System.out.println("     Number of total steps:               " + formatInt(" " + numSteps));
            System.out.println("     Number of conformations:             " + formatInt(" " + numConf));
            System.out.println("     Density vectors correlation:      "   + formatDouble(" " + correlation));
            System.out.println("     Density vectors angle:            "   + formatDouble(" " + theta));
            System.out.println("   ------------------------------------------------");
            System.out.println("\n\n");

            System.out.println("   ------------------------------------------------------");
            System.out.println("        N.   Density     Equilib.    Conformation     ");
            System.out.println("   ------------------------------------------------------");
            for (int i = 0; i < numConf; i++) {
                string = "    "  + formatInt(" " + vectorOfDensityInt[i]) +
                         "    "  + formatDouble(" " + vectorOfDensityDouble[i]) +
                         "    "  + formatDouble(" " + equilibriumVectorDouble[i]) +
                         "     " + vectorOfConformations[i];
                System.out.println(string);
            }
            System.out.println("   ------------------------------------------------------");

            System.out.println("\n\n\n");
            System.out.println("   ------------------------------------------------");
            System.out.println("                Markov Transition Matrix            ");
            System.out.println("   ------------------------------------------------");
            System.out.println("\n");
```

```
                string = "   ";
                for (int i = 0; i < numConf; i++){
                      for (int j = 0; j < numConf; j++){
                            string += formatDouble(" " + markovTransMatrixDouble[i][j]);
                      }
                      System.out.println(string);
                      string = "   ";
                }

                System.out.println("\n\n");
                System.out.println("   ************************************************  ");
                System.out.println("\n\n");

          } catch (Exception e) {
                System.out.println("\n\tError del método printResults.\n");
                System.gc();
                System.exit(0);
          }
     }// Fin del método printResults().

     /*
      * Este método da formato a las string de resultados. Por defecto
      * he escogido 5 posiciones decimales después del punto decimal.
      */

     private static String formatDouble(String string) throws Exception {
          int n = 6;
          int dotPlace = 0;
          int E_place = 0;
          String head = "";
          String body = "";
          String tail = "";
          string.trim();
          try{
                if (string.lastIndexOf("E") < 0) {
                      string = "           " + string.trim() + "000000000000000";
                      dotPlace = string.lastIndexOf(".");
```

```
                        string = string.substring(dotPlace - 2 , dotPlace + n + 1);
                  } else {
                        dotPlace = string.lastIndexOf(".");
                        E_place  = string.lastIndexOf("E");
                        head   = string.substring(0, dotPlace).trim();
                        body   = string.substring(dotPlace+1, E_place).trim();
                        tail   = string.substring(E_place + 2).trim();
                        int g  = Integer.parseInt(tail);
                        string = head + body + "000000000000000";
                        for (int i = 1; i < g; i++) {
                              string = "0" + string;
                        }
                        string = " 0." + string;
                        string = string.substring(0,n+3);
                  }
            } catch (Exception e2) {
                  System.out.println("\n\tError del método formatDouble.\n");
                  System.gc();
                  System.exit(0);
            }
            return(string);
      }// Fin del método formatDouble().



      /*
      private static String formatDouble(String string) throws Exception {
            try {
                  double value = Double.parseDouble(string);
                  DecimalFormat decFor = new DecimalFormat("#.######");
                  string = decFor.format(value);
                  string = " " + string;
            } catch(Exception e) {
                  System.out.println("\n\tError del método formatDouble.\n");
                  System.gc();
                  System.exit(0);
            }
```

```
            return(string);
      }

      private static String scienceFormat(String string) throws Exception {
            try{
                  int q = 15;
                  int length = string.trim().length();
                  while(length <= q){
                        string = " " + string;
                        length++;
                  }
            }catch (Exception e){
                  System.out.println("\n\tError del método scienceFormat.\n");
                  System.gc();
                  System.exit(0);
            }
            return(string);
      }// Fin del método scienceFormat().
      */

      /*
       * Este método da formato a las string de resultados. Por defecto
       * he escogido un ancho de 6 posiciones.
       */
      private static String formatInt(String string) throws Exception {
            int n = 0;
            string = "          " + string.trim();
            try {
                  n = string.length();
                  string = string.substring(n - 6);
            } catch (Exception e) {
                  System.out.println("\n\tError del método formatInt.\n");
                  System.gc();
                  System.exit(0);
            }
            return(string);
      }// Fin del método formatInt
```

```
      private static void howTo() {

            System.out.println(

            "\n\tMODO DE USO DEL PROGRAMA.\n\n" +

            "Se lanza desde la línea de comandos invocando la Máquina\n" +
            "Virtual de Java (JVM) pasando como argumento el nombre de\n" +
            "la clase ya compilada (sin la extensión .class). El único\n" +
            "argumento de la clase deben ser el nombre del archivo con\n" +
            "la cadena de nombres:\n\n" +

            "java MarkovChainAnalysis fileNameSequence > MarkovResults.txt\n\n" +

            "El formato de este archivo ha de ser:\n" +
            " 1) Una sola conformación por línea.\n" +
            " 2) Una sola columna.\n\n" +

            "El programa detecta automáticamente el tamaño de la columna.\n" +
            "La salida la realiza al final de todo el proceso por la STDOUT\n" +
            "de modo que se hace necesario redireccionar manualmente el flujo\n" +
            "hacia un archivo de texto (MarkovResults.txt).\n\n" +

            "\t\t\t\t\tKepa K. Burusco\n" +
            "\t\t\t\t\t[20-II-2009]\n\n");

            System.gc();
            System.exit(0);


      }// Fin del método howTo().


}// Fin de la Clase MarkovChainAnalysis
```

12.4 Statistics.exe [C-Shell Script]. Saturation Analysis for Simulated Annealing and Molecular Dynamics

```
#!/bin/csh
#
#  [26-XI-2007]
#
#*********************************************************************************************************************************
#                          Statistics for Simulated Annealing and Molecular Dynamics
#*********************************************************************************************************************************
#
#

cd ${PWD}

if ( $1 == '-h' ||  $1 == 'h' || $1 == 'help' || $1 == '-help' || $#argv == 0 ) then

cat << EOC


  [26-XI-2007]

  Este script procesa el archivo de nombres que
  sale del NameGiantCyD.class y crea 2 archivos
  de salida en el que vuelca las conformaciones
  que van apareciendo junto con su frecuencia de
  aparición en grupos de conformaciones que van
  en potencias de 2 hasta completar el total de
  las estructuras que contiene el archivo de
  entrada.

  Así pues, genera 2 archivos:

   1.- LONG.txt   que contiene toda la
                  información, y,
   2.- SHORT.txt  que es un resumen del
                  total de conformaciones.


                           Kepa K. Burusco


EOC

else
```

```
set file = ${argv[1]}

#************************************************************
# La primera extracción es especial y no entra en el bucle.
# Coge directamente la primera estructura.
#

set numLines      = 1
set numStructures = 1

echo ' '                                                                                    >> __statFULL_"$$".tmp
echo '  -. En el grupo de estructuras' "${numLines}"                                        >> __statFULL_"$$".tmp
set numConf = `cat ${file} | head -"${numStructures}" | awk '{print $1}' | sort | uniq -c | wc -l`
echo '     Conformaciones diferentes:' "${numConf}"                                         >> __statFULL_"$$".tmp
echo ' '                                                                                    >> __statFULL_"$$".tmp
cat ${file} | head -"${numStructures}" | awk '{print $1}' | sort | uniq -c | sort -n -k1 -r    >> __statFULL_"$$".tmp
echo ' '                                                                                    >> __statFULL_"$$".tmp

#************************************************************
# El bucle se encarga de todos los grupos de conformaciones que
# vamos a hacer y que son potencias de 2 pero sin sobrepasar el
# número de conformaciones total del archivo.
#

set numStructures = 2

while ( "${numStructures}" < `cat ${file} | wc -l` )

  @ numLines = `expr "${numStructures}" + 0`

  echo ' '                                                                                  >> __statFULL_"$$".tmp
  echo '  -. En el grupo de estructuras' "${numLines}"                                      >> __statFULL_"$$".tmp
  set numConf = `cat ${file} | head -"${numStructures}" | awk '{print $1}' | sort | uniq -c | wc -l`
  echo '     Conformaciones diferentes:' "${numConf}"                                       >> __statFULL_"$$".tmp
  echo ' '                                                                                  >> __statFULL_"$$".tmp
  cat ${file} | head -"${numStructures}" | awk '{print $1}' | sort | uniq -c | sort -n -k1 -r  >> __statFULL_"$$".tmp
  echo ' '                                                                                  >> __statFULL_"$$".tmp

  @ numStructures = `expr 2 '*' "${numStructures}"`

end


#************************************************************
# El último grupo de conformaciones siempre es el archivo completo
# así que simplemente aplicamos los criterios del bucle anterior a
```

```
    # este conjunto de estructuras con independencia de que sea una
    # potencia de 2.
    #

    @ numLines = `cat ${file} | wc -l`

    echo ' '                                                                                    >> __statFULL_"$$".tmp
    echo '  -. En el grupo de estructuras' "${numLines}"                                        >> __statFULL_"$$".tmp
    set numConf = `cat ${file} | awk '{print $1}' | sort | uniq -c | wc -l`
    echo '     Conformaciones diferentes:' "${numConf}"                                         >> __statFULL_"$$".tmp
    echo ' '                                                                                    >> __statFULL_"$$".tmp
    cat ${file} | awk '{print $1}' | sort | uniq -c | sort -n -k1 -r                            >> __statFULL_"$$".tmp
    echo ' '                                                                                    >> __statFULL_"$$".tmp


    #*********************************************************************
    # Este fragmento de código hace una clasificación del archivo temporal
    # __statFULL...tmp extrayendo en dos columnas el número de estructuras
    # y el número de conformaciones diferentes que hay en esas estructuras.
    # Esta información aparecerá en los dos archivos de salida que genera
    # este script al comienzo del archivo.
    #

    echo ' '                                                                                >  __estr_"$$".tmp
    echo 'Estructuras'                                | awk '{printf "%16s\n", $1}'       >> __estr_"$$".tmp
    cat  __statFULL_"$$".tmp | grep 'estructuras'     | awk '{printf "%16s\n", $7}'       >> __estr_"$$".tmp
    echo ' '                                                                                >> __estr_"$$".tmp
    echo ' '                                                                                >  __conf_"$$".tmp
    echo 'Conformaciones'                             | awk '{printf "%16s\n", $1}'       >> __conf_"$$".tmp
    cat  __statFULL_"$$".tmp | grep 'Conformaciones' | awk '{printf "%16s\n", $3}'       >> __conf_"$$".tmp
    echo ' '                                                                                >> __conf_"$$".tmp
    paste  __estr_"$$".tmp __conf_"$$".tmp                                                  >  __statSHORT_"$$".tmp

    rm -rf __estr_"$$".tmp >& /dev/null
    rm -rf __conf_"$$".tmp >& /dev/null

    #*****************************************************************************
    # Este fragmento de código genera los dos archivos de salida finales con el análisis
    # estadístico de las conformaciones. Después, todos los archivos intermedios son
    # eliminados.
    #

    @ numLines = `cat ${file} | wc -l`

    echo ' '                                                                                    >  SHORT.txt
    echo '  -. En el grupo de estructuras' "${numLines}"                                        >> SHORT.txt
```

```
    set numConf = `cat ${file} | awk '{print $1}' | sort | uniq -c | wc -l`
    echo '     Conformaciones diferentes:' "${numConf}"                                    >> SHORT.txt
    echo ' '                                                                               >> SHORT.txt
    cat ${file} | awk '{print $1}' | sort | uniq -c | sort -n -k1 -r                       >> SHORT.txt
    echo ' '                                                                               >> SHORT.txt


    cat    __statSHORT_"$$".tmp                                                            >  LONG.txt
    cat    __statFULL_"$$".tmp                                                             >> LONG.txt

    rm -rf __statSHORT_"$$".tmp >& /dev/null
    rm -rf __statFULL_"$$".tmp  >& /dev/null

endif

#*********************************************************************************************************************************************
exit 0
#*********************************************************************************************************************************************
```

12.5 dynamicLambda.exe [C-Shell Script]. Dynamic Lambda index for Molecular Dynamics

```
#!/bin/csh
#
#  [4-V-2009]
#
#*********************************************************************************************************************************************
#
#*********************************************************************************************************************************************

cd ${PWD}

if ( $1 == '-h' ||  $1 == 'h' || $1 == 'help' || $1 == '-help' || $#argv == 0 ) then

cat << EOC


  [4-V-2009]

  Este script procesa los archivos de nombres que
  salen del NameGiantCyD.class:
     nameItzi_sN.txt
     nameKepa_sN.txt
  y genera un archivo de salida en el que cita
```

```
  el numero total de conformaciones en cada DM
  leyendo paso a paso desde el comienzo del
  archivo, de una en una, hasta completar el
  total de las estructuras que contiene el
  archivo de entrada.
  Este paso lo repite para cada uno de los
  archivos de las DM individuales y para el
  archivo combinado de las trajectorias, que
  previamente genera paso a paso a partir de
  los archivos de las DM individuales.


                          Kepa K. Burusco


EOC

else

    set file = "${argv[1]}"

    #*************************************************************
    # El primer paso determina el numero total de trayectorias
    # que van a ser analizadas y el numero de conformaciones por
    # trayectoria.
    #

    set i = 1
    while ( -f ${file}_s${i}.txt )
      set numTraj = "${i}"
      @ i++
    end
    echo '  -. Se han detectado '"${numTraj}"' trayectorias'

    set numLines = `cat ${file}_s1.txt | wc -l`
    echo '  -. Se han detectado '"${numLines}"' conformaciones por trayectoria'

    #*************************************************************
    # Este fragmento de codigo genera el archivo combinado de
    # trayectorias. Para ello lee paso a paso y de una en una
    # las conformaciones de las trayectorias individuales y las
    # escribe de forma seguida en el archivo combinado.
    # Suponiendo que tuvieramos 3 trayectorias el proceso seria:
    # la conf 1 de la t1, la conf 1 de la t2, la conf 1 de la t3,
    # la conf 2 de la t1, la conf 2 de la t2, la conf 2 de la t3
    # y asi sucesivamente...
```

```
#

if ( -f  comb"$$".tmp ) then
  rm -rf comb"$$".tmp >& /dev/null
endif

set k = 1
while ( ${k} <= ${numLines} )
  set t = 1
  while ( ${t} <= ${numTraj} )
    cat ${file}_s${t}.txt | head -${k} | tail -1 >> comb"$$".tmp
    @ t++
  end
  @ k++
end

echo '  -. Se ha creado el archivo combinado de las '"${numTraj}"' trayectorias'

#*********************************************************************
# Comenzamos a crear los archivos de conformaciones para las
# trayectorias. Evalua el numero de conformaciones diferentes
# que aparecen segun avanza el proceso de DM. Para ello calcula
# cuantas conformaciones diferentes hay si se considera, desde
# el principio, la primera conformacion, luego las 2 primeras,
# luego las 3 primeras, luego las 4 primeras, y asi sucesivamente
# hasta considerar la trayectoria completa.
# Cuando la trayectoria haya explorado todo el espacio conformacional
# no aumentara el numero total de conformaciones aunque aumentemos
# el numero de pasos, y la trayectoria habra saturado.
#

set t = 1

while ( "${t}" <= "${numTraj}" )

  if ( -f  traj"${t}"."$$".tmp ) then
    rm -rf traj"${t}"."$$".tmp >& /dev/null
  endif

  echo "        - procesando la trajectoria individual traj${t}"

  echo t"${t}"                                                                    >  traj"${t}"."$$".tmp

  set k = 1
  while ( "${k}" <= "${numLines}" )
    cat ${file}_s${t}.txt | head -"${k}" | awk '{print $1}' | sort | uniq -c | wc -l       >> traj"${t}"."$$".tmp
```

```
      @ k++
    end

    echo "        - analizada  la trajectoria individual traj${t}"

    @ t++

  end

  echo '  -. Se han analizado las '"${numTraj}"' trayectorias individuales'

  #*********************************************************************
  # Esta porcion del codigo analiza el archivo combinado de trayectorias
  # para lo cual lee el archivo en pasos, de N en N, siendo N el numero
  # de trayectorias empleadas para crear el archivo combinado.
  #

  if ( -f  average"$$".tmp ) then
    rm -rf average"$$".tmp >& /dev/null
  endif
  if ( -f  counter"$$".tmp ) then
    rm -rf counter"$$".tmp >& /dev/null
  endif

  echo 'ave'                                                                    >  average"$$".tmp
  echo 'count'                                                                  >  counter"$$".tmp

  set k = 1
  set numStructures = 1
  while ( "${k}" <= "${numLines}" )
    @ numStructures = `expr "${k}" '*' "${numTraj}"`
    cat comb"$$".tmp | head -"${numStructures}" | awk '{print $1}' | sort | uniq -c | wc -l  >> average"$$".tmp
    echo "${k}"                                                                 >> counter"$$".tmp
    @ k++
  end

  echo '  -. Se ha analizado la trayectoria combinada'

  #*********************************************************************
  # Este fragmento de codigo da formato a los archivos intermedios generados
  # hasta este punto, genera el archivo final de salida, y borra los archivos
  # temporales una vez ha termiando el proceso.
  #

  cat counter"$$".tmp       | awk '{printf "%7s\n", $1}'  >   c"$$".tmp
  cat average"$$".tmp       | awk '{printf "%7s\n", $1}'  >   a"$$".tmp
```

```
    set t = 1
    while ( "${t}" <= "${numTraj}" )
      cat traj"${t}"."$$".tmp | awk '{printf "%7s\n", $1}'  >   t"${t}"."$$".tmp
      @ t++
    end

    if      ( "${numTraj}" == 1 ) then
      paste c"$$".tmp t1."$$".tmp a"$$".tmp                                          > "${file}".LAMBDA.txt
    else if ( "${numTraj}" == 2 ) then
      paste c"$$".tmp t1."$$".tmp t2."$$".tmp a"$$".tmp                              > "${file}".LAMBDA.txt
    else if ( "${numTraj}" == 3 ) then
      paste c"$$".tmp t1."$$".tmp t2."$$".tmp t3."$$".tmp a"$$".tmp                  > "${file}".LAMBDA.txt
    else if ( "${numTraj}" == 4 ) then
      paste c"$$".tmp t1."$$".tmp t2."$$".tmp t3."$$".tmp t4."$$".tmp a"$$".tmp      > "${file}".LAMBDA.txt
    else if ( "${numTraj}" == 5 ) then
      paste c"$$".tmp t1."$$".tmp t2."$$".tmp t3."$$".tmp t4."$$".tmp t5."$$".tmp a"$$".tmp  > "${file}".LAMBDA.txt
    else
      echo '  -. ERROR. Se han encontrado más de 5 trayectorias y'
      echo '     este script no puede procesar mas de 5 de modo que'
      echo '     adapte el script para este caso.'
      echo '     Terminando el proceso'
    endif

    echo '  -. Se ha creado con exito el archivo de salida.'

    #*****************************************************************
    # En este paso se borran todos los archivos temporales y se acaba
    # el proceso...
    #

    rm -rf  comb"$$".tmp     >& /dev/null
    rm -rf  traj*."$$".tmp   >& /dev/null
    rm -rf  average"$$".tmp  >& /dev/null
    rm -rf  counter"$$".tmp  >& /dev/null
    rm -rf  c"$$".tmp        >& /dev/null
    rm -rf  a"$$".tmp        >& /dev/null
    rm -rf  t*."$$".tmp      >& /dev/null

    echo '  -. Proceso terminado con exito.'

endif

#*************************************************************************************************************************************
exit 0
#*************************************************************************************************************************************
```