TESI DOCTORAL

CARACTERITZACIÓ I MILLORA DE MODELS ESTRUCTURALS DE BAIXA RESOLUCIÓ

Memòria presentada per David Piedra Garcia
per optar al grau de Doctor per la Universitat de Barcelona

Aquesta tesi doctoral ha estat realitzada sota la direcció del Dr. Xavier de la Cruz Montserrat, de la unitat de Modelatge Molecular i Bioinformàtica del Parc Científic de Barcelona.

Director
Dr. Xavier de la Cruz Montserrat

Autor
David Piedra Garcia

Índex de Continguts

ÍNDEX DE CONTINGUTS 1

AGRAÏMENTS 9

CAPÍTOL 1. INTRODUCCIÓ I OBJECTIUS 11

1.1. MÈTODES DE PREDICCIÓ ESTRUCTURAL 17
MODELAT PER HOMOLOGIA 17
THREADING O FOLD RECOGNITION 28
PREDICCIÓ *AB INITIO* O *DE NOVO* 30

1.2. QUALITAT ESTRUCTURAL 37
QUALITAT EN PREDICCIONS 38
QUALITAT EN PARTS FUNCIONALS 41

1.3. OBJECTIUS 43

1.4. REFERÈNCIES 45

CAPÍTOL 2. METODOLOGIA GENERAL 53

2.1. BASES DE DADES BIOLÒGIQUES 55
BASE DE DADES PDB: PROTEIN DATA BANK 55
BASE DE DADES CATH 57

2.2. PROGRAMES DE COMPARACIÓ ESTRUCTURAL 61
MAMMOTH 63
SSAP 63
CE 64
LGA 65
LSQ 66

2.3. PROGRAMES D´ANÀLISI ESTRUCTURAL 67
NACCESS: CÀLCUL D´ÀREA ACCESSIBLE AL SOLVENT 67
DSSP 68

SURFNET 69
CX 70
PROSA 71

2.4. MODELLER **73**

2.5. XARXES NEURONALS **75**
APRENENTATGE I FUNCIONAMENT DE LES XARXES NEURONALS 76
VALIDACIÓ CREUADA 76
PROBLEMA DEL DESBALANÇ DE CLASSES 77
MESURA DE RENDIMENT DE LES XARXES NEURONALS 77
XARXA NEURONAL UTILITZADA 78

2.6. REPRESENTACIONS GRÀFIQUES **79**
DIAGRAMES DE CAIXES O BOXPLOTS 79
VISUALITZACIÓ DE MOLÈCULES 80

2.7. REFERÈNCIES **81**

CAPÍTOL 3. IDENTIFICACIÓ DE LA FAMÍLIA ESTRUCTURAL DE PREDICCIONS *DE NOVO* **85**

3.1. INTRODUCCIÓ **87**

3.2. METODOLOGIA **89**
CONJUNT DE PREDICCIONS *DE NOVO* 89
DOMINIS SREP DE CATH 90
EINES DE COMPARACIÓ ESTRUCTURAL 90
XARXA NEURONAL 90

3.3. RESULTATS **91**
PROTOCOL D´IDENTIFICACIÓ DE FAMÍLIA ESTRUCTURAL 91
CONTRAST DEL PROTOCOL 95

3.4. DISCUSSIÓ **97**

3.5. REFERÈNCIES **99**

CAPÍTOL 4. IDENTIFICACIÓ DE ZONES CORRECTES

EN PREDICCIONS *DE NOVO* 101

4.1. INTRODUCCIÓ 103

4.2. METODOLOGIA 107
PREDICCIONS *DE NOVO* 107
REPRESENTANTS CATH 107
MÈTODES DE COMPARACIÓ ESTRUCTURAL 109
RMSD 110
GDT_TS 110
MD, SLS i SAS 110
ÀREA DE SUPERFÍCIE ACCESSIBLE I ESTRUCTURA SECUNDÀRIA 111
IDENTITAT DE SEQÜÈNCIA 112
PROSA 112

4.3. RESULTATS 115
PROTOCOL SCLQA 115
CARACTERITZACIÓ DE LES PARTS ALINEADES 116
RMSD 117
GDT_TS 121
DISTRIBUCIÓ DELS STR AL LLARG DE LA SEQÜÈNCIA 122
MILLORA DE LA QUALITAT DELS STR 124
LÍMITS DE SCLQA 126

4.4. DISCUSSIÓ 129

4.5. REFERÈNCIES 131

CAPÍTOL 5. REFINAT DE PREDICCIONS *DE NOVO* 135

5.1. INTRODUCCIÓ 137

5.2. METODOLOGIA 141
CONJUNT DE PREDICCIONS 141
ALINEAMENTS ESTRUCTURALS 141
MODELAT PER HOMOLOGIA 142
CÀLCUL D´RMSD I GDT_TS 142

5.3. RESULTATS **143**

5.4. DISCUSSIÓ **149**

5.5. REFERÈNCIES **151**

CAPÍTOL 6. EFECTE DE LES MUTACIONS SOBRE

LES CAVITATS DEL LISOZOM HUMÀ **155**

6.1. INTRODUCCIÓ **157**
CAVITATS DE PROTEÏNES I MODELS EVOLUTIUS 158

6.2. METODOLOGIA **161**
MUTANTS DEL LISOZIM HUMÀ 161
CÀLCUL DE CAVITATS 161
COMPARACIÓ DE LES CAVITATS 162
FLUCTUACIONS TÈRMIQUES ATÒMIQUES 164
FREQÜÈNCIA D´INTEREACCIONS MODIFICADES AL VOLTANT DEL PUNT DE MUTACIÓ 164

6.3. RESULTATS **167**
EFECTES GENERALS DE LES MUTACIONS 168
EFECTE DE LA LOCALITZACIÓ DE LA MUTACIÓ 170
EFECTE DE LA NATURALESA DE LA MUTACIÓ 173

6.4. DISCUSSIÓ **177**
EVOLUCIÓ I AL.LOSTERISME 178

6.5. REFERÈNCIES **181**

CAPÍTOL 7. CONSERVACIÓ DE LES CAVITATS

EN ELS MODELS PER HOMOLOGIA **187**

7.1. INTRODUCCIÓ **189**

7.2. METODOLOGIA **193**

SELECCIÓ DELS PARELLS TARGET–TEMPLATE 193
PROTOCOL DE MODELAT PER HOMOLOGIA 195
IDENTITAT DE SEQÜÈNCIA 196
CÀLCUL DE CAVITATS 196
CANVIS EN LES CAVITATS 196

7.3. RESULTATS 201
CANVI EN LA FORMA DE LES CAVITATS 201
FACTORS QUE AFECTEN A LA QUALITAT DE LA CAVITAT 213

7.4. DISCUSSIÓ 219

7.5. REFERÈNCIES 221

CAPÍTOL 8. CONCLUSIONS GENERALS 227

LLISTA DE PUBLICACIONS 231

Agraïments

"*Lo único que envidio es el saber*"
Higinio Garcia Jurado

Curiosa manera de començar a escriure uns agraïments, però com a científic crec que tinc el deure de començar per allò més general, i de fet essencial. Raonament deductiu crec que en deien. Ja quan era petit vaig sentir aquesta frase, no a la televisió ni la ràdio, i ni molt menys a una internet aleshores inexistent. La vaig sentir de boca del meu avi. Suposo que d´alguna manera una frase tant simple em va quedar gravada de forma perpètua. Així que els primers agraïments van dirigits al meu avi, per fer créixer en mi les ganes de saber, de conèixer tot el que em rodeja, per fer-me tenir inquietuds, en definitiva, per fer-me ser una mica més científic.

I suposo que després d´aquesta emotiva introducció ja puc passar als agraïments formals. En segon lloc, i per seguir amb l´ordre lògic que algú va establir un bon dia, voldria agrair als meus pares i germana per confiar en mi, per sentir-se orgullosos del que faig –encara que no sàpiguen molt bé que és això de la bioinformàtica–. En tercer lloc lògic, però primer dins el meu cor, agrair a la meva *coseta* Cristina per fer-me sentir viu, per la seva paciència, per haver pagat alguns plats que havia trencat el monstre que viu dins tot becari: la incertesa i nervis sobre el futur.

En quart lloc toca el torn a la gent de Puig-reig, en ordre alfabètic: Buixa, Maldo, Margó, Marzà, Yuste, Albert, a tots els de Gironella, als amics que van i venen, i a tots els que m´han aportat coses al llarg de la vida. Gràcies pel vòlei, partides a la consola, sortides en bici, rialles, xerrades *frikis*, gràcies també per l´anagrama de la que fou la meva primera escola (eh Cristian?); però sobretot gràcies per fer-me descobrir allò realment important de la vida, per fer-me estimar allà on visc, fer-me pertànyer a un lloc.

Agrair com no en cinquè lloc a la gent del dia a dia: al Sergi per ser la meva ànima bessona en aquesta etapa de la vida, als "Ivans" per aportar llum i rialles informàtiques, a la Rebeca per aportar maduresa als problemes –i ajudar-me amb el disseny de la portada–, al Nachete per la classe i educació que encomana, a l´Oliver per ser físic de professió i artista de vocació, a l´Agustí per aprendre a estimar el voleibol, a l´Adam per compartir el sentiment de ser de poble, al David per ensenyar-me a ensenyar, al Jose per la calma amb tocs de vainilla que desprèn, l´Agnès i l´Albert pels debats i les psico-bromes, al Manu per aquell primer mail explicant com automatitzar el *telnet*, a l´Alberto per aquella ampolla "d´aigua" a les Avellanes, al Ramon per les seves idees extravagants, a la gent del BSC pels partidets de futbol, al Jordi per aportar claredat sobre el curiós i desconegut món dels peixos, a la Montse per demostrar-me que família i ciència poden anar de la mà, a la Marga per convertir la feina administrativa i burocràtica en quelcom planer, a la Teresa per l´ajuda en el dipòsit de la tesi. Als que heu aparegut fa poc moltes gràcies per estar presents d´alguna manera o altra en aquesta etapa final, i als que possiblement em descuidi perdó per l´oblit i mil gràcies també per la part que segur que us toca.

Ja per acabar, sols em queda agrair als doctors Modesto Orozco, Josep Lluís Juan Fernández-Recio, Francisco Javier Luque, Miquel Calvo i gent de l´equip CATH per les opinions, crítiques i consells al llarg d´aquests anys, i en especial al meu director de tesi el doctor Xavier de la Cruz, per la confiança dipositada en mi, per ensenyar-me a ser científic i polir petits defectes, per les rialles i xerrades a 30º C, pels dinars a la bolera, per les barbacoes a la muntanya, i pel seu sentit de l´humor; gràcies també per les xerrades sobre futbol, sobre música i altres temes que no tenen res a veure amb la ciència; la ciència és maca però què seria d´ella si no la poguéssim barrejar amb una dosi de banalitat?

Quan vaig començar aquest viatge sovint pensava en aquest dia, tot tipus d´emocions en afloraven al meu cap. Avui puc dir que l´emotivitat queda superada per la satisfacció, la satisfacció de poder escriure una pàgina atapeïda d´agraïments i probablement quedar-me curt, d´haver tingut tanta gent al meu costat aportant-me coneixement, moments i vivències. Moltes gràcies a tots per formar part d´aquest, a vegades tortuós, viatge.

CAPÍTOL 1.

INTRODUCCIÓ I OBJECTIUS

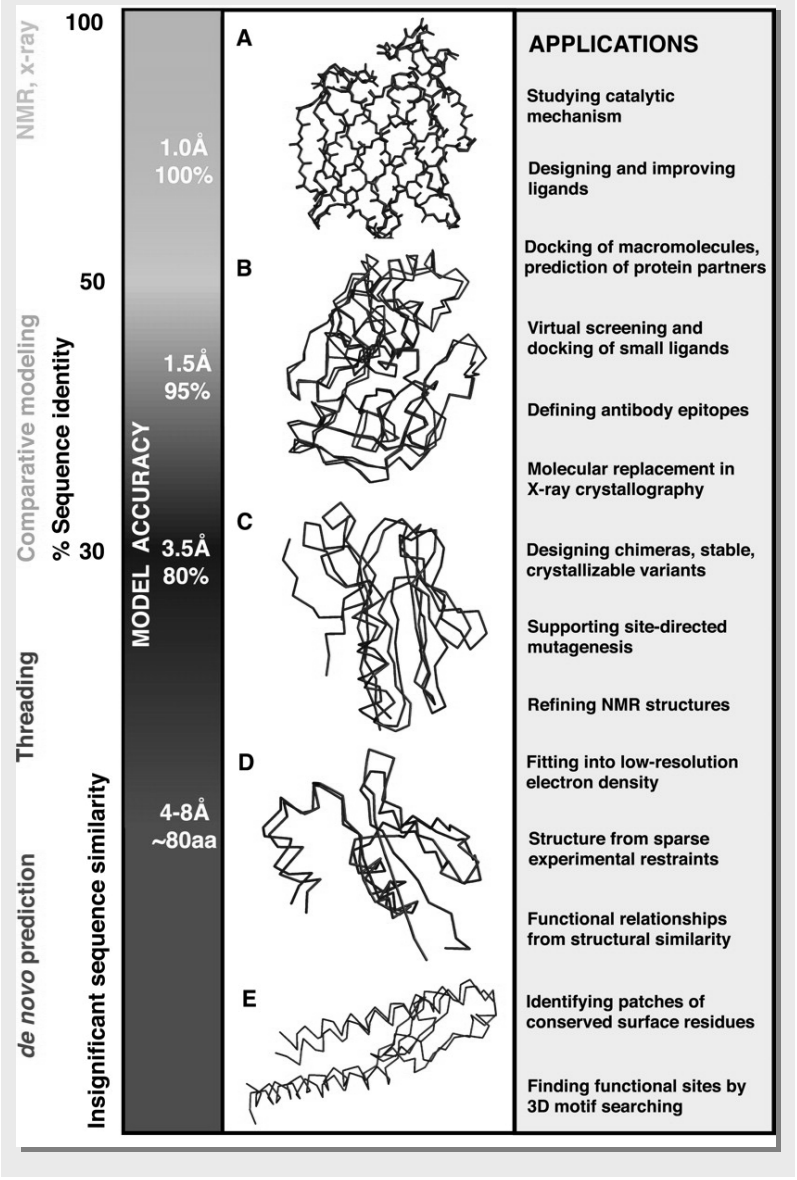
Encara avui la biologia molecular, tot i haver aconseguit fites tant importants com la seqüenciació del genoma humà, es troba davant de reptes importants. Un exemple és entendre com les proteïnes es pleguen per donar una estructura tridimensional única. No hi ha dubte que coneixem molt sobre les proteïnes: seqüència, funció, activitat, etc però encara no som capaços de predir correctament l´estructura que tindran coneixent únicament la seva seqüència. Sens dubte, el camp del plegament de les proteïnes roman com un dels enigmes de la biologia molecular, sobretot si tenim en compte que dins la cèl.lula una cadena peptídica és capaç de plegar-se en mil.lisegons.

Què ens aporta conèixer l´estructura tridimensional d´una proteïna? Per posar alguns exemples, l´estructura ens pot ajudar a determinar la funció, el mecanisme d´acció o el paper que poden jugar en determinades patologies; també pot servir-nos per dissenyar fàrmacs de forma més racional o per comprendre els efectes de les mutacions en l´activitat dels enzims.

Òbviament la predicció del plegament de proteïnes no és quelcom que hagi aparegut recentment sinó que ja té unes quantes dècades d´història. Cap a mitjans dels anys 70 amb els treballs de Levitt van aparèixer els primers intents de simular el plegament de cadenes peptídiques [1]. Des d´aleshores nombrosos estudis han estat realitzats, i nombrosos són els avenços aconseguits. En l´evolució d´aquest camp cal remarcar l´aparició dels experiments CASP el 1994 (Critical Assessment of Techniques for Protein Structure Prediction, http://predictioncenter.org/). Els experiments CASP es realitzen cada dos anys, i proporcionen una avaluació objectiva dels mètodes de predicció del moment. Consisteixen en predir un seguit de seqüències amb estructura resolta, però no pública, per la comunitat experimental; d´aquesta manera els diversos grups participants poden posar a prova els seus mètodes de predicció estructural. Per als biòlegs serveix de guia per escollir els mètodes de predicció que millor s´adjustin a les seves necessitats; per als investigadors que treballen en predicció estructural els experiments CASP serveixen per mesurar l´eficàcia de les seves tècniques. No cal dir

que el caire competitiu d´aquest experiment ha ajudat a millorar els mètodes de predicció al llarg de les diferents edicions.

Figura 1.1. Aplicació de les prediccions estructurals segons la seva qualitat.



Amb tot això, quina és la situació actual de la predicció estructural? Doncs tal i com es desprèn del darrer experiment CASP el 2006 (CASP7) [2], tot i que en alguns casos és possible fer prediccions d´estructura força fiables, és difícil obtenir resultats bons de forma sistemàtica; en altres paraules, la qualitat de les prediccions obtingudes és molt variable. Amb això apareix una pregunta: són útils aquestes prediccions? La veritat és

que no és necessari que una predicció tingui una gran qualitat per ser utilitzada; no obstant si que cal tenir present que la qualitat limitarà l´àmbit d´ús. Al següent esquema extret de [3] es pot veure com fins i tot a resolucions baixes les prediccions poden ser útils per estudiar llocs funcionals (aquest punt es tractarà al darrer capítol), o efectes de mutagènesi.

Predicció estructural en el context de la genòmica estructural

Amb l´aparició dels projectes de seqüenciació de genomes sencers el forat entre informació estructural i informació de seqüència està esdevenint cada vegada major. No cal dir que conèixer els genomes és útil, per exemple en l´establiment de relacions evolutives; no obstant això el complet enteniment i aprofitament de tota aquesta informació inevitablement requereix conèixer l´estructura tridimensional i funció de les proteïnes per les quals codifica.

Resulta evident que determinar de forma experimental l´estructura de totes aquestes seqüències seria inviable des d´un punt de vista econòmic (i moltes vegades tècnic), per tant no sembla una solució real per expandir la cobertura estructural a tot el proteoma. Com a solució alternativa apareixen els projectes de genòmica estructural. Aquests projectes, duts a terme per diversos consorcis internacionals, intenten per mitjà de metodologies experimentals i computacionals generar informació estructural de forma massiva, per tal de reduir el forat existent entre estructures i seqüències de proteïnes.

La metodologia de predicció utilitzada és el modelat per homologia, tot i que en alguns casos s´ha complementat amb mètodes *de novo* [3], per exemple per refinar determinades parts que no han estat ben modelades. El problema associat és que la qualitat del model obtingut depèn directament de l´alineament entre la proteïna a modelar i el *template*; per tant, si la identitat és baixa el risc d´obtenir models dolents és elevat, tal i com es debatrà en apartats posteriors.

Tot i que hi ha diferents aproximacions a la determinació massiva d´estructures, una típicament utilitzada és la centrada en models; aquesta es pot dividir en dues etapes [4]:

1. Etapa experimental: es resol el mínim d´estructures experimentalment. Aquestes es trien en base un seguit de criteris com pot ser interès, que presentin molts homòlegs, etc.
2. Etapa computacional: les estructures resoltes passen a utilitzar-se com a *templates* per altres seqüències homòlogues.

Actualment hi ha nombrosos projectes en funcionament, i el número d´estructures que obtenen i prediuen és força elevat. De totes maneres queda el dubte de si la qualitat és suficientment elevada com per que la comunitat científica pugui fer-ne ús.

1.1. MÈTODES DE PREDICCIÓ ESTRUCTURAL

A continuació es tractaran els mètodes de predicció estructurals utilitzats en la determinació estructural computacional: modelat per homologia, *fold recognition* (o *threading*) i mètodes *de novo* o *ab initio*.

Gran part del treball d´aquesta tesi està relacionat d´alguna manera o altra amb el primer i el darrer, és per això que seran tractats de forma més extensa.

MODELAT PER HOMOLOGIA

L´objectiu del modelat per homologia és construir un model tridimensional d´una proteïna d´estructura desconeguda (*target*) en base a la identitat de seqüència a una proteïna d´estructura coneguda (*template*). S´han de complir dues condicions: en primer lloc la similitud entre la seqüència *target* i el *template* ha de ser detectable; en segon lloc, ha de ser possible generar un alineament correcte entre les seves seqüències [5]. El modelat per homologia és possible ja que petits canvis en la seqüència d´una proteïna solen traduir-se en petits canvis en la seva estructura 3D [6].

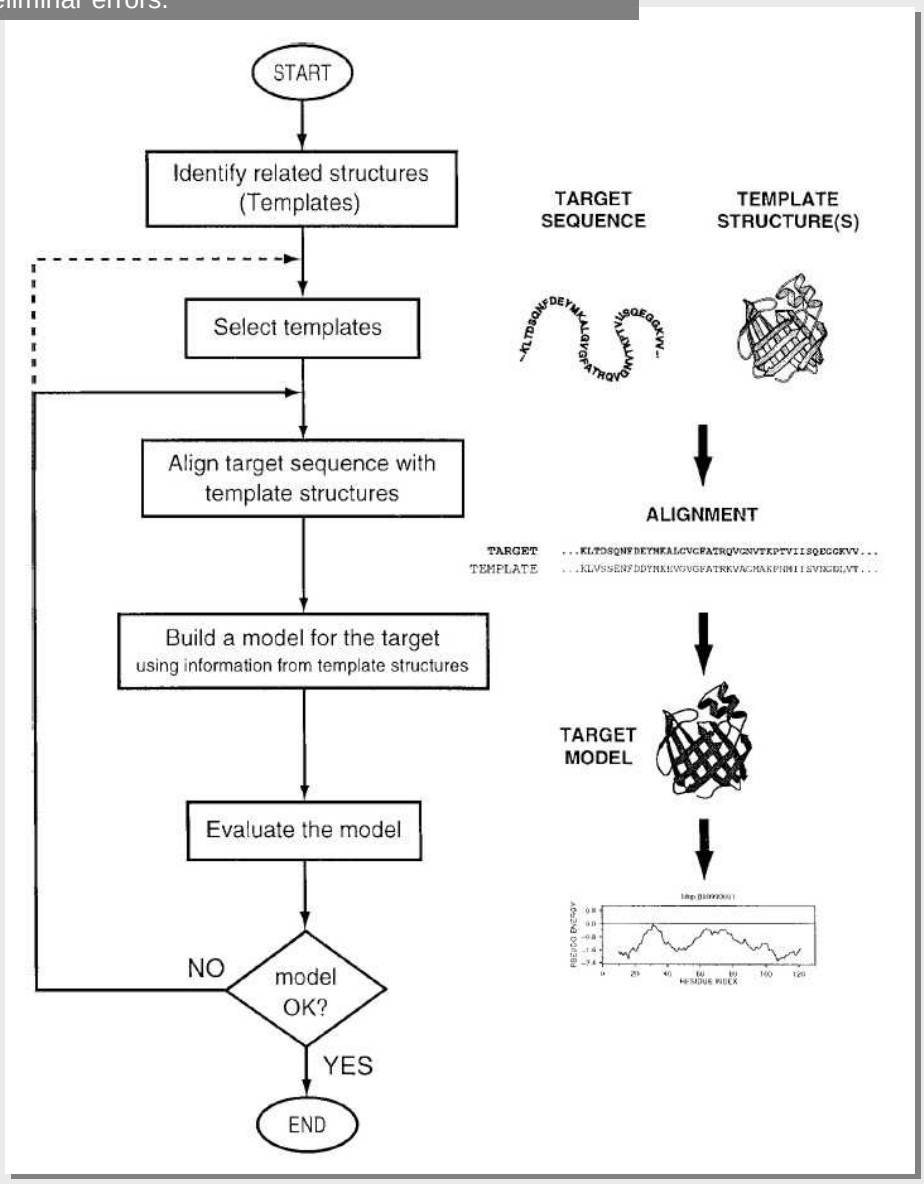
Tot i que altres camps de la predicció estructural, com la predicció *de novo*, han experimentat progressos considerables [2], avui en dia el modelat per homologia continua sent el mètode de predicció més acurat. Com es comentarà més endavant, el rang de qualitats que es poden obtenir és força ampli, no obstant això fins i tot els models menys acurats poden tenir alguna utilitat.

ETAPES DEL MODELAT PER HOMOLOGIA

Genèricament, els mètodes de modelat per homologia consten de 4 etapes: selecció del template, alineament del *target* i *template*, construcció del model, i avaluació

d´aquest. A la figura 1.2 es pot veure un diagrama del procés (extret de [5]). Cal comentar que per cadascuna de les etapes existeixen nombrosos programes, mètodes i servidors web disponibles. Existeixen també programes que reprodueixen totes les etapes; un dels més utilitzats és Modeller [7]. Al llarg d´aquesta tesi ha estat utilitzat en nombrosos estudis; a l´apartat de *Metodologia general* es pot consultar més amb detall el funcionament.

Figura 1.2. Protocol general del modelat per homologia. Essencialment hi ha tres passos: identificació dels *templates*, alineament del *template* amb la proteïna *target* i construcció del model. Addicionalment els models generats són avaluats per tal d´eliminar errors.



Selecció del template.

El pas inicial en el modelat per homologia és identificar totes les estructures de proteïnes relacionades amb la seqüència del *target* (homòlegs). D´entre aquestes estructures es podrà seleccionar la que serà utilitzada com a *template*.

El pas d´identificació d´homòlegs s´efectua comparant la seqüència *target* amb la seqüència de les proteïnes d´estructura coneguda. La comparació pot ser essencialment de tres tipus:

- Comparació seqüència-seqüència: s´utilitzen programes com BLAST [8], o FASTA [9]. D´aquesta manera es poden trobar aquelles estructures amb seqüència similar al *target*.
- Comparació seqüència-seqüència iterativa: es basa en l'ús d´alineaments de múltiples seqüències, per tal d´augmentar la capacitat de detecció d´homòlegs. Un programa típicament utilitzat per fer aquest tipus de comparacions és PSI-BLAST [10]. En aquesta aproximació es realitza una primera cerca de seqüències que ens proporciona una llista inicial d´homòlegs. Sobre la llista de homòlegs trobats s´efectua un alineament múltiple, es construeix un alineament consens i aquest s´utilitza per fer una nova cerca contra la base de dades. Aquest pas es repeteix fins que cap nou homòleg és trobat. Aquest tipus de comparació és especialment útil per trobar *templates* quan la identitat de seqüència amb el *target* és baixa (propera al 25%).
- *Threading*: donada una seqüència *target*, se li imposa l´estructura tridimensional d´un conjunt de possibles *templates*, i s´avalua si l´estructura generada és energèticament favorable. El *threading* o *fold recognition* és en si mateix un mètode de predicció, i com a tal serà discutit més endavant.

Una vegada es disposa de la llista d´homòlegs es poden seleccionar aquelles estructures que es faran servir com a *template*. Com a norma general el *template* serà aquella proteïna que comparteixi més identitat amb el *target*; no obstant en alguns

casos pot interessar tenir en compte altres consideracions com per exemple les condicions en que ha estat resolta la seva estructura (pH, temperatura), que presenti o no presenti un lligand concret, etc.

Alineament *target-template*.

Molts dels mètodes utilitzats per identificar homòlegs poden generar alineaments de seqüència; no obstant això aquests alineaments no són òptims, ja que els mètodes en que es basen estan adaptats per trobar relacions remotes entre seqüències.

Això fa que, una vegada seleccionat el *template*, sigui necessari realitzar un altre alineament de seqüència entre ell i el *target*. Aquest pas és el més important en el modelat per homologia, i el que limitarà la qualitat del model final. Només l´estructura de les zones alineades pot ser modelada en base a l´estructura del *template*, per tant com millor sigui l´alineament millor serà la zona modelada. Existeix una gran varietat de mètodes d´alineament de seqüència, molts d´ells basats en programació dinàmica. En la majoria de casos, estan basats en algorismes com Needleman i Wunsch [11]. Per exemple, el programa Modeller [5] utilitza un algorisme de programació global dinàmica amb penalitzacions per *gap* optimitzades que posiciona els *gaps* en un context estructural millor.

En casos més complicats on un *template* cobreix una part reduïda de la seqüència del *target*, és usual utilitzar varis *templates*. Altres possibilitats passen per utilitzar perfils en comptes de seqüències [12]: l´alineament es duu a terme entre un alineament múltiple que conté la seqüència *target*, i un alineament múltiple entre els diferents *templates*.

Construcció del model.

Una vegada s´ha generat l´alineament entre el *target* i el *template* es construeix el model. Existeixen diversos mètodes de construcció, amb rendiments força similars si s´utilitzen de forma òptima. De fet, altres factors com la selecció del *template*, o la

obtenció d´un bon alineament *target-template* juguen un paper més important en la qualitat final del model. Seguidament es comenten tres dels mètodes de construcció més utilitzats:

- Modelat per unió de cossos rígids: fou el primer mètode utilitzat. Es basa en la unió d´un petit nombre de cossos rígids obtinguts de l´alineament de les estructures dels *templates*. Aquest tipus de modelat és utilitzat pel programa Composer [13]: primer superposa i alinea els *templates*; en segon lloc s´amitjanen les coordenades dels Cα de les regions estructuralment conservades a les estructures dels *templates*; en tercer lloc aquestes coordenades són transferides a la zona del *target* amb la identitat de seqüència més elevada.
- Modelat per aparellament de segments: els Cα de les regions conservades entre *templates* serveixen de guia. La resta d´àtoms són afegits de manera que encaixin amb les regions conservades [14].
- Modelat per satisfacció de restriccions espacials: aquest mètode es basa en imposar un seguit de restriccions estructurals al *target*, utilitzant l´alineament amb els *templates* com a guia. En altres paraules, s´assumeix que les distàncies i els angles a les zones alineades entre *target* i *templates* seran similars. Una vegada transferides les restriccions, es duu a terme una minimització de les violacions de les restriccions [7].

Modelat de *loops*.

Les zones sense estructura secundària com els *loops* poden tenir un paper funcional molt important dins la proteïna. Per exemple els *loops* desenvolupen un paper important en la unió de substrats als centres actius (ex. unió d´ions metàl.lics a les *metal-binding proteins*); en la unió anticòs-antígen (ex. Immunoglobulines); etc.

No obstant això, la flexibilitat dels *loops*, una de les seves característiques més destacables, els converteixen en parts difícils de modelar; recordem que en el modelat

per homologia la construcció del model de la seqüència *target* es realitza en base un alineament de seqüència, i normalment les parts alineades (que solen correspondre a les menys variables) són les que es modelen millor.

Quines possibilitats existeixen alhora de modelar *loops*? És un problema difícil; de fet el modelat de *loops*, en particular *loops* grans, es pot considerar com un problema de *mini-protein-folding* [5], a causa de l´espai conformacional involucrat.

Tal i com passa en la predicció estructural de les proteïnes, hi ha diverses aproximacions utilitzades per predir la conformació dels *loops*. Majoritàriament poden ser de dos tipus:

- Mètodes *de novo*: es basen en explorar les diferents conformacions que pot adoptar el *loop*, guiant el procés per una funció d´energia o puntuació [15-17].
- Mètodes basats en bases de dades: consisteixen en buscar segments en estructures existents (homòlegs o no homòlegs) que tinguin una certa similitud al *loop* a modelar, i que encaixin geomètricament dins les zones que el delimiten. Una vegada s´ha trobat un segment candidat, és superposat i sotmès a una minimització d´energia [18-20].

Modelat de cadenes laterals.

En general, els mètodes de modelat per homologia automàtics com Modeller no són capaços de fer bones prediccions de cadenes laterals [5]; és per això que existeixen programes explícits. Aquests programes es basen en llibreries de rotàmers [21-23], que contenen informació sobre els angles de torsió de les conformacions preferides de les cadenes laterals. El problema és que a mesura que el nombre de rotàmers incrementa, es fa més difícil mostrejar totes les possibles conformacions. Existeixen alternatives per reduir l´espai conformacional, per exemple ús de llibreries de rotàmers basats en coordenades d´estructures cristal.litzades i no en angles i distàncies idealitzades [24].

FONTS D´ERROR EN EL MODELAT PER HOMOLOGIA

L´aplicació dels passos anteriors permet obtenir un model estructural per al *target*. No obstant això, les diferents aproximacions utilitzades en la seva construcció comporten un seguit d´errors que afecten a les diverses parts de l´estructura. Essencialment aquests errors poden dividir-se en cinc categories [25]:

1. Errors en les cadenes laterals: com s´ha comentat en l´apartat anterior els programes automàtics de modelat per homologia no donen resultats òptims en la predicció de cadenes laterals. Aquests errors poden ser de vital importància si tenen lloc en regions involucrades en la funció de la proteïna, com per exemple centres actius o llocs d´unió a lligand.
2. Distorsions i desplaçament en zones correctament alineades: com a conseqüència de la divergència de seqüència, les zones mal modelades poden ocasionar un decreixement de qualitat d´aquelles parts ben alienades.
3. Errors en regions sense *template*: els segments del *target* que no presenten regió equivalent en el *template* són difícils de modelar, ja que no se´ls pot imposar cap tipus de restricció geomètrica. Cal dir que regions sense alinear petites, de fins a 9 residus, en moltes ocasions poden ser predites de forma correcta, tot i que per norma general, la qualitat de les zones no alineades és baixa.
4. Errors causats per alineaments erronis: els alineaments erronis són la principal font d´error en el modelat per homologia, ja que és l´alineament el que guia la transferència d´informació estructural des del *template* cap al *target*. En molts casos és recomanable revisar o millorar els alineaments a ma.
5. Ús de *templates* incorrectes: escollir un *template* poc adequat limitarà la qualitat del model final. De fet una mala elecció en el *template* pot ser la causa d´errors en l´alineament.

AVALUACIÓ I REFINAMENT DELS MODELS

Els errors comentats en l´apartat anterior no afecten de la mateixa manera a totes les parts de la proteïna, per exemple, com s´ha anat comentant les zones alineades solen estar millor modelades que la resta. Per tal de facilitar l'ús dels models per part de la comunitat científica és important identificar les parts del model amb major qualitat. Tot i que no existeix un protocol definitiu a seguir per determinar si un model és correcte o no, hi ha diversos criteris que poden variar segons l'ús que vulguem donar al model.

Una primera aproximació a la qualitat del model pot ser fixar-nos en la identitat de seqüència entre *target* i *template*; si és inferior al 30% és molt probable que el model contingui errors importants. A part d´aquest primer criteri es poden realitzar altres comprovacions:

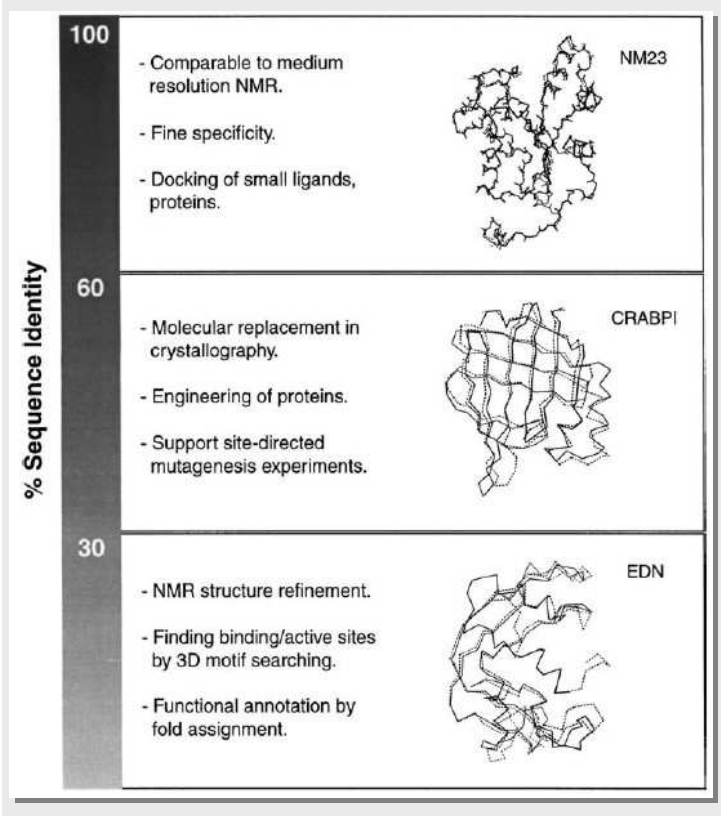
- Determinació del *fold*: una primera avaluació pot ser comprovar si el plegament del model és correcte [26]. Un model presentarà un plegament correcte si mostra una similitud alta amb el *template* que s´ha utilitzat per construir-lo (o alguna altra estructura de la mateixa família). La similitud pot buscar-se a nivell estructural, energètic, o per conservació de les parts funcionals.
- Comprovació de l´estereoquímica: existeixen programes com Procheck [27], Whatcheck [28], etc. que analitzen paràmetres geomètrics com longitud o angles d´enllaç, quiralitat, angles de torsió, etc. Com que els models estat construïts en base a estructures experimentals, els valors d´aquestes magnituds no haurien d´allunyar-se dels valors típics d´aquestes.
- Comprovació de l´entorn dels residus: mètodes basats en perfils 3D com ANOLEA [29] i potencials estadístics de força mitjana com ProsaII [30]. Aquests programes avaluen l´entorn de cada residu del model respecte a l´entorn que s´esperaria trobar en una estructura experimental.

Una vegada avaluat el model pot sotmetre´s a un refinament, per tal de millorar la qualitat de les zones de pitjor qualitat. Així per exemple és comú utilitzar programes de refinament de *loops* [31-39] i de cadenes laterals [21-23] o fins i tot dinàmiques moleculars a escales de temps llargues. De totes maneres, com s´ha comentat a l´apartat de fonts d´error, el pas limitant continua essent aconseguir bons alineaments.

APLICACIONS DELS MODELS PER HOMOLOGIA

Com s´ha vist el modelat per homologia és una forma senzilla i eficient de generar models estructurals. Aquests models poden emprar-se en diversos camps com ara disseny de mutants, estudi de funció, identificació de centres actius, disseny de fàrmacs, etc.

Figura 1.3. Aplicació de les prediccions obtingudes per modelat per homologia segons la seva qualitat.



En la figura 1.3 es poden veure algunes de les aplicacions dels models. En rangs d´identitat de seqüència baixos molts models encara presenten un plegament correcte, per tant poden ser útils per fer prediccions aproximades de funció; de fet, tal i com es veurà al darrer capítol en alguns casos tot i que globalment la resolució dels models sigui baixa les parts funcionals poden estar ben modelades. Quan la identitat de seqüència es troba entre el 30-60% els models poden ser utilitzats per predir amb més detall aspectes funcionals del centre actiu, o construir mutants. En models on la identitat és elevada, les resolucions que s´obtenen son similars a les d´estructures experimentals de baixa resolució (al voltant de 3Å); poden ser utilitzats en disseny de fàrmacs o estudi de *docking* amb lligands i amb altres proteïnes.

SITUACIÓ ACTUAL DEL MODELAT PER HOMOLOGIA

En els experiments CASP existeix una categoria de predicció emprant modelat per homologia: TBM (*Template Based Modelling*). Els resultats permeten tenir una idea de la situació actual, eficiència i rendiment d´aquests mètodes.

En la categoria TBM del CASP7 (2006) es van presentar un total de 187 grups, aportant 15717 prediccions de 108 dominis estructurals. D´aquestes prediccions un subconjunt és avaluat dins la categoria d´alta precisió, on es miren característiques com orientació de cadenes laterals, o aplicabilitat en recanvi molecular (ús de models per determinar les fases dels mapes de densitat electrònics d´estructures resoltes per raig X). Pel que fa als resultats d´aquest darrer CASP, els punts més importants es resumeixen a continuació [40]:

- Hi ha mètodes automàtics que comencen a donar resultats prometedors, similars als millors resultats obtinguts per grups que utilitzen mètodes amb intervenció manual.
- Habitualment el modelat per homologia té una limitació: donat un model format a partir d´un *template*, el template tendeix a ser més proper a l´estructura experimental que el propi model. Hi ha força mètodes que han

aconseguit superar aquest punt. L´explicació és que l'ús de múltiples *templates* comença a utilitzar-se de forma efectiva.

- El punt anterior per altra banda demostra que utilitzant únicament un *template* és complicat obtenir millores, amb el que es pot concloure que la identificació de bons *templates*, i l´obtenció de bons alineaments continua essent la limitació principal del modelat per homologia.
- En alguns casos, tot i que la qualitat global dels models és bona, si ens fixem en zones funcionals aquesta qualitat decreix. És important millorar aquest aspecte, ja que la utilitat d´una predicció passa per un bon modelat de les zones funcionals.
- Pel que fa als models d´alta precisió, hi ha algun grup que comença a obtenir bons resultats en els criteris utilitzats [41].

De totes maneres, tot i que els resultats del darrer experiment CASP són força optimistes i demostren que els mètodes de modelat per homologia donen cada vegada millors resultats, la realitat és que encara hi ha molts punts pendents de millora o solució. Així per exemple existeix una limitació intrínseca del mètode: la necessitat de disposar d´homòlegs amb estructura resolta. Un altre punt a tenir en compte és que ara per ara, utilitzant els mètodes actuals, és difícil modelar de forma acurada detalls d´alta precisió com són la rotació de cadenes laterals.

THREADING O FOLD RECOGNITION

El nom *threading* fou emprat per primer cop el 1992 [42], tot i que els primers estudis són anteriors [43, 44]. La idea bàsica del *threading* o *fold recognition* és imposar a una seqüència *target* l´estructura d´una col.lecció de proteïnes o *templates*, i analitzar quin dels models resultants és més favorable. Com a criteri d´anàlisi s´utilitza una funció d´energia.

Es pot considerar que el *threading* o *fold recognition* es troba entre el modelat per homologia i la predicció *de novo*, ja que comparteix característiques d´ambdós; així per exemple tal i com succeeix amb el modelat per homologia el *threading* es basa en trobar un seguit de *templates* amb qui alinear una seqüència sense estructura. De forma similar als mètodes *de* novo, els mètodes de *threading* utilitzen funcions d´energia per descartar si un plegament és o no correcte. En situacions on només es disposa d´estructura amb baixa homologia respecte la seqüència a modelar, el *threading* és una bona opció.

ETAPES TÍPIQUES DEL *FOLD RECOGNITION* O *THREADING*

Essencialment consisteix en alinear una seqüència amb un conjunt d´estructures conegudes o *templates*, i avaluar quina estructura dóna una puntuació més elevada. Posteriorment es pot transferir la informació espacial del *template* a la seqüència problema per mitjà de l´alineament *target–template*, de forma similar a com es feia en el modelat per homologia.

Típicament podem parlar de 4 etapes:

Representació del target.

Existeixen diverses formes de representar la seqüència *target*; una de les formes més habituals és la representació de la seqüència com a perfil P, on cada element *Pj* és un vector amb una distribució de probabilitats dels 20 aminoàcids en la posició *j*. Aquest

perfil es construeix típicament a partir d´alineaments múltiples de seqüències no redundants [45]. Alguns mètodes de *threading* incorporen a més a cada posició informació sobre estructura secundària [46] o altra informació derivada de la seqüència com pot ser l´accessibilitat al solvent.

Representació dels *templates*.

Pel que fa a les estructures utilitzades com a *templates*, les coordenades 3D es redueixen a representacions que poden ser més abstractes. La principal classificació dels mètodes de *threading* està lligada al tipus de representació que es fa de l´estructura dels *templates*. Essencialment els algorismes existents són de dos classes:

- Reducció a perfil 1D: un exemple senzill d´aquest tipus de representació seria etiquetar cada aminoàcid de l´estructura segons si es troba enterrat o exposat a la superfície. Altres perfils més elaborats poden tenir en compte estructura secundària, o informació evolutiva. El principal avantatge és que treballar amb aquest tipus de representacions és computacionalment ràpid.
- Ús d´informació 3D: els perfils inclouen informació estructural sobre els residus, per exemple distàncies interatòmiques respecte àtoms veïns, o directament les coordenades dels Cα o Cβ [44, 47]. D´aquesta manera s´incorpora informació sobre interaccions entre residus. Existeixen mètodes més complexos que permeten capturar regularitats com pot ser la hidrofobicitat [48]. El problema d´aquests sistemes de representació és que al ser més complexes també són computacionalment més costosos.

Alineament de la seqüència al *target*.

La representació dels *templates* determinarà el procés d´obtenció de l´alineament entre aquests i el *target*. Així per exemple, quan el *template* es representa com a perfil 1D, l´alineament es pot construir emprant algorismes de programació dinàmica com Needleman i Wunsch [11]; el resultat és un alineament òptim des del punt de vista de la funció de puntuació. En el cas dels *templates* representats amb informació 3D, els

mètodes per generar els alineaments esdevenen més complexos, i es necessiten un altre tipus d´algorismes com per exemple Montecarlo, per als que és impossible garantir que proporcionin un alineament òptim.

Avaluació dels alineaments.

La majoria de mètodes de *threading* utilitzen funcions determinades empíricament per anàlisi estadístic d´estructures de proteïnes presents al PDB [49], és per això que sovint es coneix aquestes funcions com a potencials empírics. Aquest tipus de funcions d´energia són molt utilitzades en la predicció *de novo*. En l´apartat corresponent a aquests mètodes de predicció es discutirà la base teòrica.

SITUACIÓ ACTUAL DEL *THREADING*

Els mètodes de *threading* tenen com a principal avantatge que poden ser emprats en situacions on la baixa identitat de seqüència entre *target* i *template* limita l'ús dels mètodes com el modelat per homologia.

Tal i com succeeix amb el modelat per homologia, els mètodes de *threading* s´avaluen durant els experiments CASP. Diversos grups han aconseguit bons resultats [50-52], no obstant això la realitat és que el número de prediccions correctes en general pot variar dramàticament segons el cas que intentem predir.

La situació del *threading* doncs és similar a la del modelat per homologia: s´estan aconseguint avenços, i així ho demostren els experiments CASP [53]; no obstant això aquesta millora continua essent massa dependent del creixement de les bases de dades d´estructures.

PREDICCIÓ *AB INITIO* O *DE NOVO*

L´objectiu original de la predicció *ab initio* o *de novo* és generar l´estructura tridimensional a partir d´una seqüència, emprant únicament les lleis bàsiques de la

física. A diferència dels mètodes de modelat per homologia o *threading* en principi no es necessita cap tipus d´informació sobre homòlegs o proteïnes relacionades, per tant en situacions on és impossible utilitzar aquests mètodes pot ser útil recórrer als mètodes *de novo*.

El principi en que es basen els mètodes *de novo/ab initio* per a l´obtenció de l´estructura del *target* és assumir que quan una proteïna es plega, ho fa adoptant la conformació de més baixa energia lliure [54]. No obstant això, l´aplicació d´aquest principi dista molt de ser fàcil, ja que tal i com es postula a la paradoxa de Levinthal [55], el número de conformacions geomètricament possibles d´una cadena proteica és enorme: una proteïna de 100 aminoàcids, amb tres conformacions possibles per aminoàcid, tindria 3^{100} possibles estats conformacionals, dels quals un seria el natiu (o un energèticament similar [56]).

Això fa que resoldre el problema del plegament passi per dos necessitats:

- Desenvolupar un mètode eficient per mostrejar l´espai conformacional.
- Desenvolupar una funció d´energia, el mínim absolut o mínims globals de la qual corresponguin a una conformació similar a l´estructura de l´estat natiu de la proteïna.

MOSTREIG DE L´ESPAI CONFORMACIONAL

Com s´ha comentat, l´espai conformacional d´una cadena de 100 residus amb tres graus de llibertat és enormement extens. Mostrejar-lo tot seria computacionalment impossible. És per això que en les tècniques *de novo* es recorre a la reducció de complexitat.

Essencialment hi ha dos tipus de reducció de complexitat: els mètodes basats en xarxes, i els mètodes no basats en xarxes.

- Mètodes basats en xarxes: representen els àtoms de les cadenes peptídiques dins d´una reixa tridimensional, de tal manera que els àtoms se situen als punts que defineixen l´entramat. Essencialment el que s´aconsegueix d´aquesta manera és discretitzar les coordenades dels àtoms. El principal avantatge és la simplicitat analítica i computacional que ofereixen. En contraposició trobem que les restriccions geomètriques fan difícil representar elements d´estructura secundària, com per exemple les hèlix. Aquest punt es pot solucionar reduint la distància de les cel.les de la reixa, tot i que això comportaria un increment en el cost computacional [57, 58].
- Mètodes no basats en reixes: es basen en l´eliminació de graus de llibertat de les cadenes peptídiques. Això es pot aconseguir simplificant les cadenes laterals (anul.lant-les, o conservant només Cβ per exemple), o limitant els valors possibles dels angles phi/psi de la cadena principal (discretitzant els valors als que es donen amb més freqüència). Existeixen diversos estudis basats en discretització d´angles amb resultats interessants [58, 59].

FUNCIONS D´ENERGIA

Tenint en compte que la conformació adoptada per una proteïna és la de menor energia lliure, o una propera a ella [56], és necessari disposar d´una funció que ens permeti avaluar el terme energètic de les conformacions generades en el pas anterior.

Una funció d´energia hauria de englobar tots els termes que contribueixen de forma significativa al plegament de les proteïnes, tals com interaccions electrostàtiques, efecte hidrofòbic, etc. No obstant això, tot i que existeixen diverses teories sobre la termodinàmica del plegament de proteïnes [60, 61], encara no es disposa d´una funció completament adequada, a causa de la complexitat del problema. Les aproximacions emprades pels diferents autors han donat lloc a un ventall de funcions d´energia que es poden classificar de forma genèrica en dues classes: funcions d´energia físiques i funcions d´energia estadístiques.

- Funcions d´energia físiques: intenten recollir tots els factors que contribueixen en el plegament de la proteïna: ponts d´hidròg en, efecte hidrofòbic, etc. El seu objectiu és reproduir, a partir de termes físics, les característiques estructurals principals de les estructures de les proteïnes. Per exemple, de les funcions que inclouen un terme corresponent a l´efecte hidrofòbic (principal força que guia el plegament d´una proteïna [62]) s´espera que reprodueixin el patró de localització dels aminoàcids: els hidrofòbics al cor de la proteïna, i els hidrofílics a la superfície de la mateixa.
- Funcions d´energia estadístiques: contenen una descripció funcional de les característiques principals de les estructures de les proteïnes, i representen la síntesi de milers d´observacions experimentals. Per exemple, aquestes funcions puntuen el patró de contactes entre un residu i els seus veïns, en funció del que s´assembla al comportament habitual de dit residu en les proteïnes d´estructura coneguda. La suma de les puntuacions de cadascun dels residus dóna una idea de fins quin punt l´estructura presenta característiques d´estructura nativa. L'ús d´aquestes funcions ha anat acompanyat de certa polèmica de caire conceptual, ja que no queda del tot clar si és correcte relacionar-les amb l´energia física del sistema [63]. Al marge d´aquesta polèmica, cal assenyalar que les funcions d´energia estadístiques han donat bons resultats.

S´ha demostrat que tant les funcions d´energia estadístiques com físiques poden ser útils en l´evaluació del plegament de proteïnes [64], i de fet amb el temps possiblement apareixeran mètodes mixtes.

ROSETTA

Rosetta és un programa de predicció estructural *de novo* creat per David Baker [65]. Va ser utilitzat per primer cop en el CASP3 (1998), i des d´aleshores és dels mètodes que millors puntuacions ha obtingut al llarg dels experiments CASP [66, 67].

Tot i ser un mètode *ab initio*, Rosetta utilitza informació estructural existent en bases de dades, així com una funció d´energia basada en potencials estadístics com les comentades a l´apartat anterior. Essencialment el funcionament del programa consta de dos passos [65]:

1. Es trenca la seqüència en fragments de 9 residus. Per cadascun dels fragments es construeix una llibreria que contindrà fragments de seqüència similar, extrets d´estructures presents al PDB. L´assumpció darrera aquest pas és que la distribució de conformacions d´un pèptid de 9 residus és aproximada a la distribució que adopta en les estructures conegudes.
2. Es construeixen les estructures terciàries utilitzant una simulació per Monte Carlo [65], mostrejant l´espai conformacional dels nonapèptids emprant les llibreries de fragments del pas anterior. Les conformacions obtingudes són avaluades amb una funció d´energia que té en compte: a) termes dependents de seqüència que representen hidrofobicitat, accessibilitat i interaccions específiques com les electrostàtiques o ponts disulfur; i b) termes independents de seqüència que representen empaquetament d´hèlix alfa i fulles beta.

L'ús de llibreries de fragments dóna resultats força bons, i des del primer experiment CASP on fou utilitzat (1998) per David Baker ha estat emprat per diversos grups. De totes maneres, com es comentarà en seccions posteriors la fiabilitat dels mètodes *de novo* en general és encara reduïda, i en molts casos està limitada a proteïnes de mida inferior als 100 residus.

APLICACIÓ DELS MODELS *DE NOVO*

Dels mètodes de predicció estructural existents el que millors resultats dóna és el modelat per homologia. Tal i com s´ha vist però, el rendiment i ús d´aquest mètode està lligat a l´existència de proteïnes amb estructura homòlogues a la seqüència

problema; per tant si aquesta condició no es dóna la utilitat del modelat per homologia decreix. És en aquest context on la predicció *de novo* pot ser utilitzada.

Per exemple, en els projectes de seqüenciació de genomes és molt habitual que aparegui un gran nombre de seqüències sense homologia amb proteïnes d´estructura/funció coneguda [68]; en aquests casos les tècniques de predicció *de novo* poden ser útils per generar models aproximats; models que poden ser utilitzats per inferir informació sobre la família estructural o funcional [69].

També en casos on el modelat per homologia sigui possible, els mètodes *de novo* poden ser un complement útil; per exemple per predir zones de baixa homologia, com *loops* o regions variables. De fet, en els darrers CASP s´està posant de manifest que mètodes de predicció híbrids donen bons resultats globals [70].

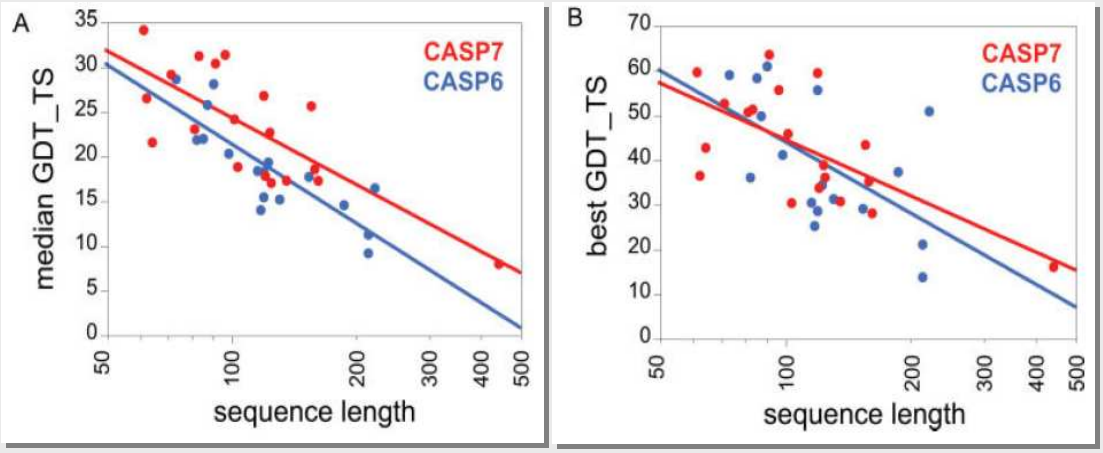
SITUACIÓ ACTUAL DE LA PREDICCIÓ *DE NOVO*

De la mateixa manera que succeeix amb el modelat per homologia i el *threading*, en els experiments CASP també s´avalua el rendiment dels mètodes de predicció *de novo*. Tot i que inicialment aquests mètodes s´havien basat en aproximacions purament físiques, amb la introducció el 1997 del mètode Rosetta [65], basat en la unió de fragments extrets de llibreries, cada cop són més els grups que utilitzen tècniques d´aquest tipus.

Pel que fa a l´eficiència d´aquests mètodes, tot i que és difícil comparar resultats entre les diferents edicions de CASP (els criteris de puntuació, mida de les proteïnes a predir per exemple poden variar), la tendència dels mètodes *de novo* és millorar: cada vegada hi ha més grups que obtenen prediccions més acurades, i de forma més automatitzada, tot i que continua essent impossible predir el plegament d´una seqüència de forma sistemàtica. Això queda de manifest si comparem les dues darreres edicions de CASP [67]: tal i com es pot apreciar a la figura 1.4A la puntuació GDT_TS mitjana de les prediccions tendeix a ser major a CASP7 que a CASP6, fet que

suggereix que un major nombre de grups són capaços d´obtenir millors resultats. No obstant, si ens fixem en la puntuació de les millors prediccions (figura 1.4B), no s´observa una millora respecte el CASP6.

Figura 1.4. Millora en les prediccions de novo a CASP7 respecte CASP6. Al panell A es mostren les puntuacions mitjanes GDT_TS respecte la mida de les seqüències a CASP6 (blau) i CASP7 (vermell). Al panell B es mostren les puntuacions de les millors prediccions a CASP6 (blau) i CASP7 (vermell).



Un altre punt a tenir en compte és que sembla ser que les millores als experiments CASP estan lligades a dos factors: increment de llibreries de fragments i dades experimentals, i increment de la potència computacional. Òbviament a mesura que les bases de dades estructurals s´enriqueixin i la potència computacional augmenti s´aniran obtenint més bons resultats.

1.2. QUALITAT ESTRUCTURAL

Els darrers experiments CASP han mostrat que tot i el ràpid progrés en la predicció estructural, és difícil obtenir de forma sistemàtica prediccions de bona qualitat. El ventall de qualitats dels models generats és força ampli, i per tant es fa completament necessari disposar de mètodes que permetin decidir si un model és o no és bo.

El principal motiu per conèixer la qualitat d´un model està lligat a la seva aplicació, tal i com s´ha vist en els models per homologia. També cal tenir en compte que cada vegada més científics d´àmbits allunyats de la predicció estructural fan ús de models, per tant es fa necessari que disposin de mesures de qualitat que els ajudin a decidir sobre el seu possible ús.

Tradicionalment, els mètodes d´assignació de qualitat estructural van aparèixer per avaluar estructures experimentals generades per raig X o RMN. Essencialment aquests mètodes es fixen en paràmetres covalents com angles d´enllaç, angles torsionals, o distàncies, i paràmetres no covalents com exposició dels àtoms al solvent i patrons de contactes.

Els paràmetres que defineixen la qualitat d´una estructura experimental són extrapolables a les prediccions, és per això que molts dels mètodes anteriors han passat a emprar-se en el camp de la predicció estructural. Així per exemple programes de modelat per homologia com Modeller recomanen com a part del seu protocol avaluar els models generats amb programes com Prochek [27] o Prosa [30]. A part dels mètodes ja existents, n´han aparegut de nous; un exemple és PCons o ProQ [71] com es veurà tot seguit.

QUALITAT EN PREDICCIONS

En casos com el modelat per homologia és fàcil tenir una idea de la qualitat aproximada que pot tenir un model únicament fixant-nos en la identitat de seqüència entre *target* i *template*. Com ja s´ha comentat si la identitat és alta (p.e. superior al 60%) podem estar força segurs que la qualitat del model serà bona. En el cas de la predicció *de novo* i per *threading* no es pot utilitzar aquesta informació per dir si un model serà o no correcte, ja que són mètodes independents de seqüència (si més no des d´un punt de vista formal), de fet com s´ha comentat tant els mètodes *de novo* com els de *threading* solen utilitzar-se quan la seqüència *target* no disposa d´homòlegs propers i per tant no pot ser aplicat el modelat per homologia.

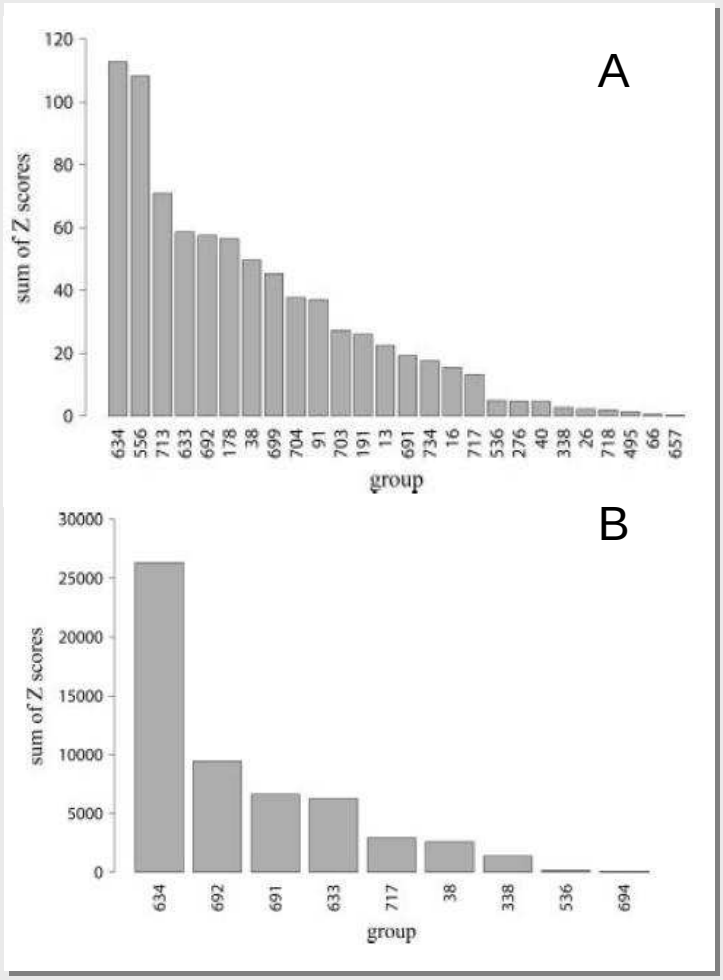
Com s´ha vist, poder disposar d´eines per avaluar la qualitat d´un model és indispensable per tenir una idea d´àmbit en el qual pot ser utilitzat. És en aquest context que en darrer experiment CASP s´ha introduït una nova categoria relacionada amb l´assignació de qualitat als models. En aquesta categoria com molt bé es pot deduir del nom, els diversos grups participants realitzen prediccions de qualitat de tots els models generats; aquestes mesures de qualitat seran comparades amb els valors establerts una vegada es coneguin les estructures reals que s´han predit. Les estimacions de qualitat són a nivell global, i a nivell de residu. L´interès de calibrar mètodes que permetin determinar la qualitat local d´una predicció cau en que poden existir prediccions globalment dolentes, però que en canvi presentin zones funcionals o d´interès de major qualitat, i viceversa. A més, definir aquelles parts més ben modelades d´una predicció pot servir per posteriors refinaments de la mateixa. Aquests dos aspectes són dos dels punts que han inspirat aquesta tesi, tal i com es podrà veure als objectius i amb més profunditat a llarg dels diversos capítols.

Avui en dia, ProsaII [30] i Verify3D [72] són dos dels programes més utilitzats per assignar qualitat local, i han estat emprats amb èxit en l´anàlisi de la qualitat de les prediccions aportades pels diversos grups en els experiments CASP. A part d´aquests

se n´han desenvolupat d´altres. Uns dels més prometedors, i que millors resultats han donat en el darrer CASP són PCons [71] i ProQ [73].

PCons és un mètode consens, que utilitza un conjunt de models. Essencialment PCons funciona superposant tots els models a avaluar i buscant patrons recurrents. Aleshores prediu la qualitat assignant una puntuació a cada model, en base a la similitud estructural que presenti respecte el conjunt. Permet obtenir una mesura de qualitat global i local. En la categoria d´assignació de qualitat del darrer CASP ha obtingut els millors resultats tant en la predicció global com en la local, tal i com es pot veure en les següents gràfiques.

Figura 1.5. Puntuacions globals per als grups participants en la categoria d´assignació de qualitat global (A) i qualitat local (B). El grup d´Arne Elofsson amb el programa PCons és el que obté millors puntuacions tant en la categoria de qualitat global com local.



Pel que fa a ProQ, utilitza una combinació de característiques estructurals per predir l´estructura global. Aquestes característiques: contactes àtom-àtom, contactes residu-residu, àrea d´exposició al solvent i informació sobre estructura secundària, són utilitzades com entrada d´una xarxa neuronal entrenada per predir qualitat.

El problema és que tant PCons com ProQ, i en general la majoria de mètodes no són utilitzables sobre models individuals, sinó que necessiten un conjunt de prediccions. Seria interessant poder disposar de mètodes que es poguessin utilitzar sobre estructures individuals, emprant únicament les coordenades dels àtoms. Al darrer CASP han aparegut mètodes d´aquest tipus, tot i que el rendiment ara per ara no és massa bo. De totes maneres no deixen de ser prometedors.

QUALITAT EN PARTS FUNCIONALS

Tradicionalment els estudis d´anàlisi de qualitat en prediccions s´ha centrat en aspectes globals. Tal i com s´ha vist al punt anterior, en el darrer experiment CASP s´ha posat de manifest la necessitat d´analitzar la qualitat no únicament des d´un punt de vista global, sinó també des d´un punt de vista local. El motiu és que un anàlisi global no és suficient per aportar detall sobre la qualitat de parts de la proteïna que poden tenir rellevància biològica, com per exemple cavitats, centres actius o punts d´unió a altres molècules.

No cal dir que la importància d´aquestes parts és enorme, ja que són les responsables de la funció de la proteïna; per exemple, s´ha demostrat que normalment les cavitats més grans de les proteïnes són les que contenen els centres actius [74]. Disposar de mètodes per mesurar la qualitat local de les prediccions estructurals és completament necessari, ja que de forma anàloga al que succeeix amb la qualitat global, la qualitat local també determinarà l´àmbit d´aplicacío de la predicció. Per exemple, si ens centrem en un centre actiu, segons la qualitat amb que estigui predit podrem utilitzar-lo en disseny de fàrmacs.

Malauradament existeixen pocs treballs relacionats amb la qualitat de les parts funcionals. Alguns dels estudis més rellevants han estat els realitzats per De Weese i Moult, que han aportat informació sobre com es preserva la qualitat de les cavitats en els models per homologia [75]. Altres estudis realitzats per R. Sanchez i A. Sali han aportat llum també sobre aspectes relacionats amb la conservació de cavitats en models per homologia, però centrant-se sobretot en la conservació de contactes amb el substrat. El problema és que tot i que aquests estudis han contribuït positivament en el tema de la qualitat local de les prediccions, han estat realitzats utilitzant pocs models, i emprant poques variables.

REFINAT

Com s´ha vist a l´apartat de predicció estructural tot i els avenços en els mètodes *de novo*, *threading* o de modelat per homologia, és encara impossible obtenir sistemàticament prediccions de bona qualitat. Cada cop més s´esta plantejant com a alternativa el desenvolupament de mètodes de refinament; d´aquesta manera la predicció estructural es pot dividir en dos punts: obtenció de prediccions de baixa resolució, i refinament d´aquestes. El principal punt a favor d´aquesta estratègia és que la primera etapa es pot dur a terme sense massa cost computacional; en la segona etapa en canvi es poden emprar mètodes computacionalment més costosos, ja que l´espai conformacional a explorar és menor.

S´han desenvolupat diverses estratègies, per exemple utilitzant dinàmiques moleculars és possible distingir entre prediccions d´estructura propera a la nativa i prediccions amb plegament incorrecte [76]; altres estudis han refinat models per homologia utilitzant programes de modelat de *loops* [31-35], o dinàmiques moleculars a escales llargues [77].

En qualsevol cas, resulta cada cop més evident que una estratègia de refinat passa abans per la identificació de les zones en les prediccions en base a la seva qualitat, per procedir a la seva millora [2]. Novament això posa de manifest la necessitat de ser capaços d´analitzar la qualitat local de les prediccions.

1.3. OBJECTIUS

Aquesta tesi ha estat desenvolupada en el context de la qualitat estructural i essencialment tracta dos punts: i) identificació de zones de qualitat en prediccions *de novo* i refinament d´aquestes, i ii) estudi de la qualitat de parts funcionals en els models per homologia.

i) En el primer bloc de la tesi ens hem proposat identificar les zones de millor qualitat d´un conjunt de prediccions *de novo*, i utilitzar-les per millorar les prediccions inicials. Aquest bloc consta de tres capítols:

- Capítol 3: Identificació de la família estructural d´un conjunt de prediccions *de novo* emprant xarxes neuronals i mètodes de comparació estructural.
- Capítol 4: Comparació estructural de les prediccions *de novo* amb membres de la seva família, i posterior extracció de les parts de millor qualitat.
- Capítol 5: Ús de les parts de millor qualitat per refinar les prediccions inicials emprant tècniques de modelat per homologia.

ii) En el segon bloc de la tesi ens hem proposat estudiar elements funcionals de les proteïnes com són les cavitats. Aquest bloc consta de dos capítols:

- Capítol 6: en aquest capítol s´ha estudiat com afecten les mutacions a les cavitats en les proteïnes natives.
- Capítol 7: en aquest darrer capítol s´ha abordat el tema de la conservació de la qualitat de zones funcionals (cavitats) en models per homologia.

1.4. REFERÈNCIES

1. Levitt, M. and A. Warshel, *Computer simulation of protein folding.* Nature, 1975. **253**(5494): p. 694-8.
2. Cozzetto, D., et al., *Assessment of predictions in the model quality assessment category.* Proteins, 2007. **69 Suppl 8**: p. 175-83.
3. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
4. Mirkovic, N., et al., *Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization.* Proteins, 2007. **66**(4): p. 766-77.
5. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.
6. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* Embo J, 1986. **5**(4): p. 823-6.
7. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
8. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
9. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.
10. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
11. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
12. Marti-Renom, M.A., M.S. Madhusudhan, and A. Sali, *Alignment of protein sequences by their profiles.* Protein Sci, 2004. **13**(4): p. 1071-87.

13. Sutcliffe, M.J., et al., *Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures.* Protein Eng, 1987. **1**(5): p. 377-84.

14. Unger, R., et al., *A 3D building blocks approach to analyzing and predicting structure of proteins.* Proteins, 1989. **5**(4): p. 355-73.

15. Bruccoleri, R.E. and M. Karplus, *Prediction of the folding of short polypeptide segments by uniform conformational sampling.* Biopolymers, 1987. **26**(1): p. 137-68.

16. Fine, R.M., et al., *Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations.* Proteins, 1986. **1**(4): p. 342-62.

17. Moult, J. and M.N. James, *An algorithm for determining the conformation of polypeptide segments in proteins by systematic search.* Proteins, 1986. **1**(2): p. 146-63.

18. Chothia, C. and A.M. Lesk, *Canonical structures for the hypervariable regions of immunoglobulins.* J Mol Biol, 1987. **196**(4): p. 901-17.

19. Greer, J., *Model for haptoglobin heavy chain based upon structural homology.* Proc Natl Acad Sci U S A, 1980. **77**(6): p. 3393-7.

20. Harbury, P.B., B. Tidor, and P.S. Kim, *Repacking protein cores with backbone freedom: structure prediction for coiled coils.* Proc Natl Acad Sci U S A, 1995. **92**(18): p. 8408-12.

21. Bower, M.J., F.E. Cohen, and R.L. Dunbrack, Jr., *Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool.* J Mol Biol, 1997. **267**(5): p. 1268-82.

22. Dunbrack, R.L., Jr. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* J Mol Biol, 1993. **230**(2): p. 543-74.

23. Ponder, J.W. and F.M. Richards, *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J Mol Biol, 1987. **193**(4): p. 775-91.

24. Xiang, Z. and B. Honig, *Extending the accuracy limits of prediction for side-chain conformations.* J Mol Biol, 2001. **311**(2): p. 421-30.

25. Sanchez, R. and A. Sali, *Evaluation of comparative protein structure modeling by MODELLER-3.* Proteins, 1997. **Suppl 1**: p. 50-8.

26. Sanchez, R. and A. Sali, *Large-scale protein structure modeling of the Saccharomyces cerevisiae genome.* Proc Natl Acad Sci U S A, 1998. **95**(23): p. 13597-602.

27. Laskowski, R.A., et al., *PROCHECK: a program to check the stereochemical quality of protein structures.* J Appl Crystallogr, 1993. **26**: p. 283-291.

28. Hooft, R.W., et al., *Errors in protein structures.* Nature, 1996. **381**(6580): p. 272.

29. Melo, F., et al., *ANOLEA: a www server to assess protein structures.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 187-90.

30. Sippl, M.J., *Recognition of errors in three-dimensional structures of proteins.* Proteins, 1993. **17**(4): p. 355-62.

31. Fiser, A., R.K. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Sci, 2000. **9**(9): p. 1753-73.

32. Rapp, C.S. and R.A. Friesner, *Prediction of loop geometries using a generalized born model of solvation effects.* Proteins, 1999. **35**(2): p. 173-83.

33. Xiang, Z., C.S. Soto, and B. Honig, *Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7432-7.

34. Zheng, Q. and D.J. Kyle, *Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings.* Proteins, 1996. **24**(2): p. 209-17.

35. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins, 2004. **55**(2): p. 351-67.

36. Li, W., Z. Liu, and L. Lai, *Protein loops on structurally similar scaffolds: database and conformational analysis.* Biopolymers, 1999. **49**(6): p. 481-95.

37. Wojcik, J., J.P. Mornon, and J. Chomilier, *New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification.* J Mol Biol, 1999. **289**(5): p. 1469-90.

38. Fidelis, K., et al., *Comparison of systematic search and database methods for constructing segments of protein structure.* Protein Eng, 1994. **7**(8): p. 953-60.

39. van Vlijmen, H.W. and M. Karplus, *PDB-based protein loop prediction: parameters for selection and methods for optimization.* J Mol Biol, 1997. **267**(4): p. 975-1001.

40. Kopp, J., et al., *Assessment of CASP7 predictions for template-based modeling targets.* Proteins, 2007. **69 Suppl 8**: p. 38-56.

41. Read, R.J. and G. Chavali, *Assessment of CASP7 predictions in the high accuracy template-based modeling category.* Proteins, 2007. **69 Suppl 8**: p. 27-37.

42. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.

43. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.

44. Hendlich, M., et al., *Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force.* J Mol Biol, 1990. **216**(1): p. 167-80.

45. Bienkowska, J. and R. Lathrop, *Threading Algorithms.* Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, 2005.

46. McGuffin, L.J. and D.T. Jones, *Improvement of the GenTHREADER method for genomic fold recognition.* Bioinformatics, 2003. **19**(7): p. 874-81.

47. Jones, D. and J. Thornton, *Protein fold recognition.* J Comput Aided Mol Des, 1993. **7**(4): p. 439-56.

48. Huang, E.S., et al., *Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations.* J Mol Biol, 1996. **257**(3): p. 716-25.

49. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

50. Skolnick, J. and D. Kihara, *Defrosting the frozen approximation: PROSPECTOR--a new approach to threading.* Proteins, 2001. **42**(3): p. 319-31.

51. Kelley, L.A., R.M. MacCallum, and M.J. Sternberg, *Enhanced genome annotation using structural profiles in the program 3D-PSSM.* J Mol Biol, 2000. **299**(2): p. 499-520.

52. Jones, D.T., *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.* J Mol Biol, 1999. **287**(4): p. 797-815.

53. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Curr Opin Struct Biol, 2005. **15**(3): p. 285-9.

54. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.

55. Levinthal, C., *ARE THERE PATHWAYS FOR PROTEIN FOLDING?* Journal de Chimie Physique, 1968. **65**(1): p. 2.

56. Shortle, D., K.T. Simons, and D. Baker, *Clustering of low-energy conformations near the native structures of small proteins.* Proc Natl Acad Sci U S A, 1998. **95**(19): p. 11158-62.

57. Kihara, D., et al., *TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10125-30.

58. Park, B.H. and M. Levitt, *The complexity and accuracy of discrete state models of protein structure.* J Mol Biol, 1995. **249**(2): p. 493-507.

59. de la Cruz, X.F., M.W. Mahoney, and B. Lee, *Discrete representations of the protein C alpha chain.* Fold Des, 1997. **2**(4): p. 223-34.

60. Ptitsyn, O.B., *Molten globule and protein folding.* Adv Protein Chem, 1995. **47**: p. 83-229.

61. Ptitsyn, O.B., et al., *Evidence for a molten globule state as a general intermediate in protein folding.* FEBS Lett, 1990. **262**(1): p. 20-4.

62. Baldwin, R.L., *Protein folding from 1961 to 1982.* Nat Struct Biol, 1999. **6**(9): p. 814-7.

63. Thomas, P.D. and K.A. Dill, *Statistical potentials extracted from protein structures: how accurate are they?* J Mol Biol, 1996. **257**(2): p. 457-69.

64. Lazaridis, T. and M. Karplus, *Effective energy functions for protein structure prediction.* Curr Opin Struct Biol, 2000. **10**(2): p. 139-45.

65. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

66. Das, R., et al., *Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home.* Proteins, 2007. **69 Suppl 8**: p. 118-28.

67. Jauch, R., et al., *Assessment of CASP7 structure predictions for template free targets.* Proteins, 2007. **69 Suppl 8**: p. 57-67.

68. Rychlewski, L., B. Zhang, and A. Godzik, *Fold and function predictions for Mycoplasma genitalium proteins.* Fold Des, 1998. **3**(4): p. 229-38.

69. Fetrow, J.S., A. Godzik, and J. Skolnick, *Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity.* J Mol Biol, 1998. **282**(4): p. 703-11.

70. Vincent, J.J., et al., *Assessment of CASP6 predictions for new and nearly new fold targets.* Proteins, 2005. **61 Suppl 7**: p. 67-83.

71. Lundstrom, J., et al., *Pcons: a neural-network-based consensus predictor that improves fold recognition.* Protein Sci, 2001. **10**(11): p. 2354-62.

72. Eisenberg, D., R. Luthy, and J.U. Bowie, *VERIFY3D: assessment of protein models with three-dimensional profiles.* Methods Enzymol, 1997. **277**: p. 396-404.

73. Wallner, B. and A. Elofsson, *Can correct protein models be identified?* Protein Sci, 2003. **12**(5): p. 1073-86.

74. Laskowski, R.A., et al., *Protein clefts in molecular recognition and function.* Protein Sci, 1996. **5**(12): p. 2438-52.

75. DeWeese-Scott, C. and J. Moult, *Molecular modeling of protein function regions.* Proteins, 2004. **55**(4): p. 942-61.

76. Lee, M.R., D. Baker, and P.A. Kollman, *2.1 and 1.8 A average C(alpha) RMSD structure predictions on two small proteins, HP-36 and s15.* J Am Chem Soc, 2001. **123**(6): p. 1040-6.

77. Fan, H. and A.E. Mark, *Refinement of homology-based protein structures by molecular dynamics simulation techniques.* Protein Sci, 2004. **13**(1): p. 211-20.

CAPÍTOL 2.

METODOLOGIA GENERAL

2.1. BASES DE DADES BIOLÒGIQUES

Les bases de dades biològiques s´han convertit en un instrument important per ajudar als científics a comprendre i explicar fenòmens biològics que van des de l´estructura biomolecular i interacció fins el metabolisme complet dels microorganismes, passant per la comprensió de l´evolució de les espècies. Aquest coneixement facilita la lluita contra malalties, ajuda en el desenvolupament de medicaments, i en el descobriment de les relacions bàsiques entre les espècies.

Tot seguit es comenten dues bases de dades àmpliament utilitzades en els treballs que conformen aquesta tesi doctoral: Protein Data Bank –base de dades d´estructura – i CATH –base de dades de classificació de dominis estructurals.

BASE DE DADES PDB: PROTEIN DATA BANK

El PDB (Protein Data Bank) és un banc de dades d´estructures tridimensionals de proteïnes i àcids nucleic, resoltes de forma experimental, majoritàriament per Ressonància Magnètica Nuclear (RMN) o per difracció de Raig X. La informació que conté és essencialment coordenades cartesianes, graus d´ocupació i factors de temperatura de tots els àtoms d´aquestes estructures, tot i que poden trobar-se referències a la seqüència, mètode d´obtenció, o altres referències bibliogràfiques.

És una base de dades de domini públic. A nivell estructural es pot considerar la base de dades a partir de la qual deriven la resta; per exemple bases de dades de classificació estructural de proteïnes (com veurem tot seguit), etc.

NOTACIONS HISTÒRIQUES

El Protein Data Bank va ser fundat l´any 1971 pels doctors Edgar Meyer i Walter Hamilton del Brookhaven National Laboratory, i el 1988 la gestió fou transferida al

Research Collaboratory for Structural Bioinformatics (RCSB). La Universitat de Rutgers és la seu principal i actualment està dirigit per Helen M. Berman.

CREIXEMENT DE LA BASE DE DADES

Quan es va fundar, PDB disposava únicament de 7 estructures de proteïnes. Des d´aleshores ha experimentat un creixement exponencial en el número d´estructures.

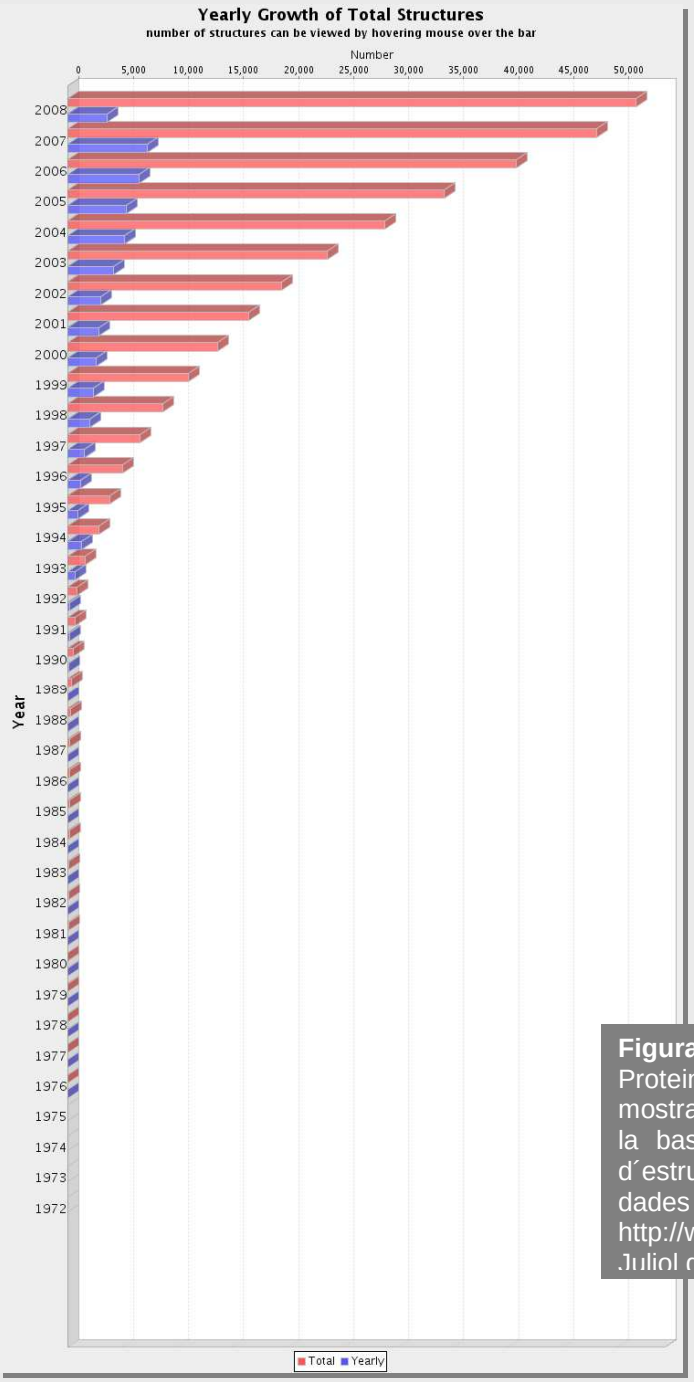


Figura 5.1. Creixement anual del Protein Data Bank. En vermell es mostra el total d´estructures presents a la base de dades; en blau el total d´estructures afegides cada any. Les dades han estat tretes de http://www.rcsb.org/pdb, data 2 de Juliol de 2008.

A 2 de Juliol del 2008 hi ha un total de 51663 estructures (incloent proteïnes i àcids nucleics). A la següent taula es pot veure de forma més detallada.

Taula 2.1. Número d´estructures al PDB a data de 2 de Juliol de 2008

	Proteines	Àcids Nucleics	Complexes Proteïna/AN	Altres	Total
Raig X	41065	1055	1874	24	44018
RMN	6410	809	138	7	7364
Micro. electrònica	124	11	47	0	182
Altres	89	4	4	2	99
Total	47688	1879	2063	33	51663

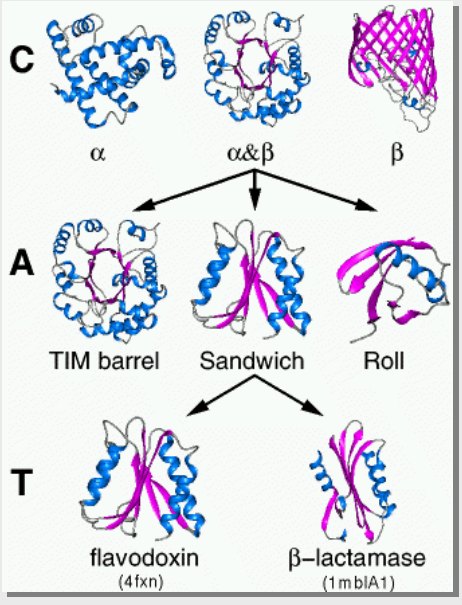
És important remarcar que no s´inclouen models teòrics en la base de dades.

Per una informació més detallada es pot visitar la pàgina web http://www.rcsb.org/pdb, on es comenta el funcionament i organització de la base de dades, les actualitzacions, el format de les dades, etc.

BASE DE DADES CATH

La base de dades CATH és una classificació jeràrquica dels dominis de les estructures emmagatzemades al Protein Data Bank [1]. Inclou estructures resoltes per RMN, i estructures resoltes per raig X sempre que la resolució sigui millor a 4Å. De la mateixa manera que el PDB no inclou models, com tampoc proteïnes amb percentatge de carbonis alfa superior al 30% respecte el total d´àtoms.

Figura 2.2. Classificació jeràrquica de la base de dades CATH. Cada nivell es ramifica en subnivells. Font: http://www.cathdb.info.



Els dominis proteics són classificats, emprant una combinació de tècniques automàtiques i manuals, en 4 nivells principals: Classe (C), Arquitectura (A), Topologia (T) i Superfamília d´homologia (H) [2].

ASSIGNACIÓ DE DOMINIS

Per tal de dividir aquelles estructures de proteïnes multidomini en els dominis constituents es combinen mètodes automàtics i manuals. Per exemple, si una determinada cadena proteica presenta una identitat de seqüència i similitud estructural elevades amb alguna cadena ja present a la base de dades (ex. Identitats de seqüència superior a 80%, o puntuació SSAP [3] superior a 80) aleshores els límits del domini són assignats de forma automàtica, prenent els de la cadena existent. En qualsevol altre cas els límits dels dominis són establerts de forma manual, basant-se en l´anàlisi de resultats derivats d´un conjunt d´algorismes que inclouen mètodes estructurals (CATHEDRAL [4], SSAP [3], DETECTIVE [5], PUU [6], DOMAK [7]), mètodes de seqüència (Perfils HMMs) i literatura rellevant.

NIVELLS CATH

Nivell de Classe (C): la classe es determina segons la composició de l´estructura secundària. Existeixen tres classes principals: alfa, beta i alfa-beta. Aquesta darrera inclou estructures amb alternança alfa-beta i alfa+beta [8]. Existeix una quarta classe que inclou dominis que presenten un contingut baix en estructura secundària.

Nivell d´Arquitectura (A): descriu la forma global del domini en base a les orientacions dels elements d´estructura secundària, sense considerar la connectivitat entre ells. El nivell d´arquitectura s´assigna de forma manual emprant una descripció senzilla de l´ordenament dels elements d´estructura secundària (ex. *Tim barrel, sandwich*).

Nivell de Topologia (T): agrupa estructures amb el mateix plegament, entenent per plegament la disposició i connectivitat dels elements d´estructura secundària.

L´assignació de Topologia es duu a terme emprant l´algorisme SSAP [3] i CATHEDRAL [4]. Aquelles estructures amb una puntuació SSAP superior a 70, on al menys un 60% de la proteïna més gran s´aparella amb la més petita es considera que pertanyen al mateix nivell T.

Nivell d´Homologia (H): aquest nivell agrupa dominis que possiblement comparteixin un ancestre comú, i per tant poden ser considerats com homòlegs. La similitud s´identifica o bé per una identitat de seqüència alta, o bé per una puntuació SSAP elevada. Més concretament, les estructures s´agrupen a un mateix nivell H si satisfan un dels següents criteris:

- Identitat de seqüència ≥ 35%, solapament ≥ 60% (estructura major respecte la menor).
- Puntuació SSAP ≥ 80, identitat de seqüència ≥ 20%, solapament ≥ 60% (estructura major respecte la menor).
- Puntuació SSAP ≥ 70, solapament ≥ 60% (estructura major respecte la menor), i dominis que tenen funcions relacionades, segons la literatura i la base de dades Pfam [9].
- Similitud significant en comparacions HMM-HMM i HMM-seqüència utilitzant SAM [10], HMMER (http://hhmer.wustl.edu) i PRC (http://supfam.org/PRC).

Nivells de seqüència (S,O,L,I,D): els dominis que pertanyen a un grup H són agrupats en diferents subnivells segons les identitats de seqüència i solapament, tal i com es pot veure a la següent taula:

Taula 2.2. Definició dels nivells de seqüència a CATH

Nivell	Identitat de seq.	Solapament
S	35%	80%
O	60%	80%
L	95%	80%
I	100%	80%
D		

Les identitats de seqüència i solapaments utilitzats per agrupar s´obtenen de la implementació de l´algorisme Needleman-Wunsch [11], utilitzant una penalització de *gap* de 3. El percentatge de la identitat de seqüència és calculat com (100 * nº residus idèntics / longitud de la seqüència més curta); el solapament com (100 * nº residus alineats / longitud de la seqüència més curta).

El nivell D únicament actua com a comptador dins de cada família I, i s´afegeix a la classificació per assegurar que cada domini únicament presenta un únic identificador CATHSOLID.

2.2. PROGRAMES DE COMPARACIÓ ESTRUCTURAL

L´alineament estructural és un tipus d´alineament basat en la comparació de la forma. Aquests alineaments intenten establir equivalències entre dos o més estructures de proteïnes basant-se en la seva forma i conformació tridimensional.

Cal diferenciar-los de les simples superposicions estructurals (veure RMSD), on es coneixen a priori els residus equivalents en les dues estructures. En els alineaments estructurals no és necessari conèixer les posicions equivalents, el programa utilitzat s´encarrega d´establir-les.

Quina utilitat pot tenir comparar estructures de proteïnes? Doncs la resposta és que la utilitat és molt àmplia; alguns dels àmbits on la comparació estructural pot tenir un paper rellevant es resumeixen en els següents punts:

- Inferència de relacions evolutives entre proteïnes amb divergència de seqüència important, on els mètodes clàssics de comparació de seqüència no són útils.
- Determinació de l´impacte estructural (i funcional) de les mutacions de residus d´una proteïna.
- Determinació de la funció proteica per comparació amb proteïnes existents ja caracteritzades.
- Verificació de prediccions estructurals.
- Estudis estructurals generals, tal i com es veurà en els propers capítols d´aquesta tesi.

A continuació s´expliquen els mètodes de comparació estructurals utilitzats al llarg d´aquesta aquesta tesi. Abans però es descriu l´RMSD, variable àmpliament utilitzada en la comparació d´estructures.

L´RMSD (de l´anglès Root Mean Square Deviation) no és un mètode de comparació estructural pròpiament dit, sinó que és una variable que mesura la distància mitjana entre els àtoms de dues estructures superposades. Es calcula utilitzant la següent fórmula:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

On δ és la distància entre els N parells d´àtoms equivalents (per exemple carbonis alfa). Tenint en compte que treballa amb distàncies, les unitats de l´RMSD seran Å. Cal tenir present que per calcular l´RMSD és necessari conèixer les equivalències entre els residus a comparar, per tal de poder calcular les distàncies entre ells. Veurem com aquest requisit no és necessari en els mètodes de comparació estructural tractats a continuació, on el mateix algorisme ja és capaç de trobar l´equivalència de residus entre les dues estructures a comparar. Per poder fer el càlcul de l´RMSD cal trobar la superposició de les dues estructures que en minimitzi el valor; això es pot aconseguir utilitzant l´algorisme Kabsch [12].

L´RMSD és una variable que depèn de la quantitat de punts comparats; així per exemple obtenir un RMSD de 4Å quan comparem dues estructures de 10 residus equival a dir que aquestes són molt diferents; en canvi si obtenim un RMSD de 4Å quan comparem dues estructures de 100 residus podem dir que són similars. Aquesta dependència es pot corregir emprant una normalització de l´RMSD respecte una mida concreta, per exemple 100 residus. Aquesta normalització en concret, ha estat utilitzada àmpliament al llarg de la tesi. L´$RMSD_{100}$ s´expressa com: RMSD / [1 + 0.5ln(N/100)], on RMSD és l´RMSD estàndard, i N és el número de punts utilitzats en el seu càlcul [13].

MAMMOTH

MAMMOTH (Matching Molecular Models Obtained from Theoretical) és un mètode d´alineament estructural ràpid i totalment independent de seqüència, que únicament té en compte les coordenades dels carbonis alfa i evita qualsevol tipus de referència a informació de seqüència o mapes de contacte. Per tal de reduir la complexitat del problema treballa emprant una aproximació heurística, de forma similar a altres mètodes existents: primer troba l´alineament estructural que proporciona la similitud local òptima de la cadena principal de la proteïna, i tot seguit intenta trobar el màxim subconjunt de residus per sota d´una distància de tal predefinida. A nivell més detallat, MAMMOTH trenca la proteïna en conjunts d´heptapèptids, i els compara amb els heptapèptids de la proteïna amb que es vol comparar. Es calcula una puntuació de similitud entre heptapèptids emprant URMS (Unit-vector Root Mean Square). Aquestes puntuacions s´emmagatzemen en una matriu de similituds, i per mitja de programació dinàmica es calcula l´alineament de residus òptim. L´algorisme complet es pot consultar a l´article original [14].

Existeix una versió que permet fer alineaments estructurals múltiples anomenada MAMMOTH-mult. Està basada en el mateix algorisme que MAMMOTH. Tant la versió senzilla com la versió d´alineament múltiple poder ser executades des de la web: http://ub.cbm.uam.es/mammoth.

SSAP

El mètode SSAP (Sequential Structure Alignment Program) utilitza doble programació dinàmica per generar un alineament estructural basat en vectors de distància àtom-àtom. En comptes d´utilitzar els carbonis alfa típicament utilitzats en programes de comparació estructural, SSAP utilitza carbonis beta (excepte per la glicina), d´aquesta manera es tenen en compte els estats rotamèrics de cada residu, així com també la seva localització al llarg de la cadena proteica.

SSAP funciona construint, per cada proteïna, un seguit de vectors de distàncies inter-residu entre cada residu i els residus veïns no contigus. Es construeixen aleshores un conjunt de matrius que contenen les diferències dels vectors veïns per cada parell de residus equivalents. La programació dinàmica aplicada a cada matriu resultant determina una sèrie d´alineaments locals òptims que són afegits a una matriu "resum" a la que se li aplica de nou programació dinàmica per determinar l´alineament estructural global.

L´algorisme detallat es pot consultar a l´article original [3], i el programa pot ser executat des del servidor http://www.cathdb.info.

CE

El mètode CE (Combinatorial Extension) trenca les estructures a comparar en fragments, normalment de 8 residus. Els fragments d´una proteïna es van alineant combinatòriament amb fragments de l´altra, donant lloc als "parells de fragments alineats" (o AFPs, de l´anglès Aligned Fragment Pairs); la combinació que dóna una similitud més alta es manté, i s´expandeix afegint un nou AFPs. Dels nous parells afegits es mantindrà aquell que doni una similitud més elevada i així successivament fins a tenir una trajectòria òptima que representi l´alineament final.

Tot i que inicialment la similitud entre fragments es determinava en base a la superposició estructural i distàncies inter-residu, s´han acabat per afegir altres propietats locals com estructura secundària, exposició al solvent, patrons de ponts d´hidrògen i angles díedres.

L´algorisme detallat es pot consultar a l´article original [15], i el programa pot executar-se o descarregar-se de la pàgina http://cl.sdsc.edu/ce.html.

LGA

El programa LGA (Local-Global Alignment) és un mètode de comparació estructural que té en compte l´estructura global i local de les proteïnes.

El programa funciona comparant distàncies entre estructures de proteïnes per a segments locals i estructura global. La funció de puntuació consta de fet de dos components: LCS (Longest Continous Segment) i GDT (Global Distance Test). L´algoritme LCS identifica les regions locals de les proteïnes a comparar on els residus tenen una similitud estructural calculada en RMSD per sota un llindar. L´algorisme GDT complementa els resultats obtinguts amb LCS buscant el conjunt més gran (no necessàriament continu) de residus equivalents que es troben per sota uns llindars de distància concrets.

En el càlcul GDT, la cerca per la superposició òptima entre dues estructures es realitza com es detalla en els següents passos:

1. Càlcul de la superposició òptima entre els àtoms de les estructures.
2. Determinació dels parells d`àtoms alineats per sota d´un llindar concret (típicament 1Å, 2Å, 4Å i 8Å).
3. Obtenció d´una nova superposició òptima utilitzant només aquells parells d´àtoms determinats al punt anterior.
4. Repetició pas 2 i 3 fins que no hi hagi canvis observats en la llista de parells d´àtoms utilitzats durant dues iteracions. El número de parells alineats serà el percentatge d`àtoms que estan per sota del llindar.
5. La puntuació final correspon a la mitjana entre els % de parells alineats per cada llindar després d´haver seguit els passos anteriors.

L´algorisme detallat es pot veure a l´article original [16], i el programa pot executar-se des de la pàgina web http://predictioncenter.llnl.gov/local/lga.

LSQ

LSQ és un programa de comparació estructural senzill [17] que combina l´alineament de seqüència i el càlcul de l´RMSD.

Funciona de la següent manera: assumint que el conjunt de parells de residus equivalents entre dues estructures pot ser determinat del corresponent alineament de seqüència [11], l´alineament estructural té lloc en dos passos:

1. Es determina el vector de translació que situa els centres de masses de les dues estructures en un mateix punt.
2. Es calcula la matriu de rotació que minimitza l´RMSD entre les dues estructures, tenint en compte únicament carbonis Alfa [12].

El programa dóna un valor d´RMSD. Per tal de tenir mesures de puntuacions similars a les obtingudes amb els altres mètodes, el codi font del programa ha estat lleugerament modificat per incorporar les funcions de puntuació suggerides a [18]: $S = 1.37 + (1.16*L-15.1)^{0.5} - R)$, i $Ss = 2.1 + S/1.8$, on R és l´RMSD i L el número de residus equivalents.

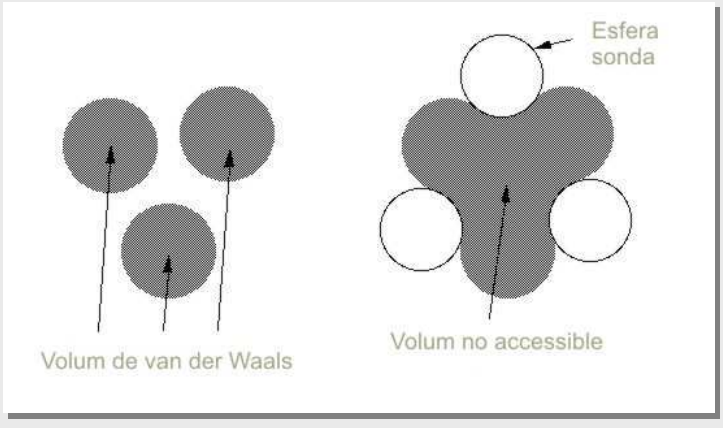
El programa pot ser descarregat http://geometry.molmovdb.org/files/lsqrms/.

2.3. PROGRAMES D´ANÀLISI ESTRUCTURAL

NACCESS: CÀLCUL D´ÀREA ACCESSIBLE AL SOLVENT

L`àrea de superfície accessible (ASA, de l´anglès Accessible Surface Area) és l`àrea de la superfície d´una biomolècula que es troba en contacte amb el solvent. Habitualment l´ASA es calcula en $Å^2$, unitat de mesura estàndard en biologia molecular. L´ASA fou descrita per primer cop per Lee i Richards el 1971 [19], i típicament es calcula tallant la proteïna en capes, i analitzant el perímetre d´àtoms accessibles al solvent. Conceptualment s´assimila a utilitzar una esfera o sonda d´un radi particular, i fer-la rodar per la superfície de la molècula; d´aquesta manera es pot comptabilitzar la superfície dels àtoms o dels residus que està en contacte amb el solvent. La mida de la sonda té efecte sobre la superfície accessible observada, així per exemple si s´utilitza una sonda de radi petit es detectaran més detalls de la superfície. Una mida típica és 1.4Å, que és aproximadament el radi de la molècula d´aigua.

Figura 2.3. Esquema conceptual del càlcul de superfície accessible atòmica.



Naccess és un programa que s´utilitza per calcular l´ASA en macromolècules; utilitza el mètode de Lee i Richards; per defecte el radi de la sonda utilitzada és 1.4Å.

L´algorisme es pot consultar als articles originals de Lee i Richards, i el programa es pot descarregar de forma gratuïta a http://www.bioinf.manchester.ac.uk/naccess.

DSSP

L´algorisme DSSP és un mètode estàndard per assignar estructura secundària als aminoàcids d´una proteïna, donades les coordenades cartesianes dels seus àtoms. L´assignació es basa en els patrons de ponts d´hidrògen. Aquest mètode utilitza una definició de pont d´hidrògen purament electrostàtica:

$$E = q_1 q_2 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} * 332\,\mathrm{kcal/mol}$$

En aquesta expressió, q_1 pren valors de 0.42 pel carboni del carbonil, i -0.42 per l´oxigen del carbonil; q_2 a la seva vegada pren valors de 0.20 pel nitrogen de l´amida, i de -0.20 per l´hidrogen. En base a aquesta fórmula, s´estableix que un pont d´hidrògen existeix si E és inferior a -0.5 kcal/mol. Segons el patró de ponts d´hidrògen és possible assignar l´estructura secundària als aminoàcids de l´estructura [20].

DSSP defineix vuit tipus d´estructura secundària, depenent del patró de ponts d´hidrògen:

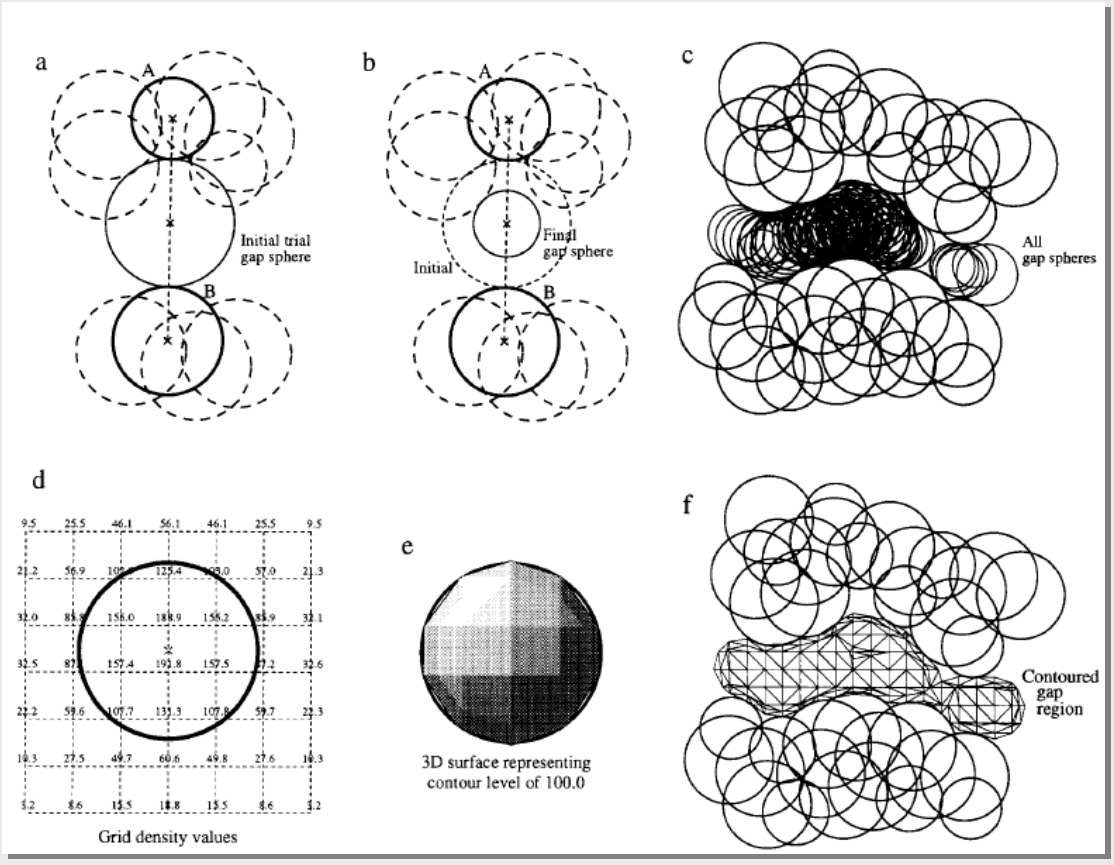
- G = hèlix 3_{10}. Volta cada 3 residus.
- H = hèlix alfa. Volta cada 4 residus.
- I = hèlix pi. Volta cada 5 residus.
- T = gir amb pot d´hidrògen. Implica 3, 4 o 5 residus.
- E = fulla beta paral.lela/anti-paralel.la. Implica mínim 2 residus.
- B = residu aïllat a un pont beta.
- S = assignació no basada en pont d´hidrogen.

SURFNET

Surfnet [21] és un programa que permet definir les cavitats existents en una proteïna. Donada una estructura tridimensional d´una proteïna, o bé un complex proteic, el programa determina les regions de la superfície que presenten una depressió, i que per tant són candidates a albergar un lloc d´unió a substrat o altres proteïnes, un centre de regulació al.lostèric, etc.

L´algorisme del Surfnet es pot resumir en els següents passos:

Figura 2.4. Representació dels passos que segueix SURFNET per calcular les cavitats. Font: http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html.



- Es situa una esfera entre un parell d´àtoms, de diàmetre tal que sigui tangent a ambdós (a).

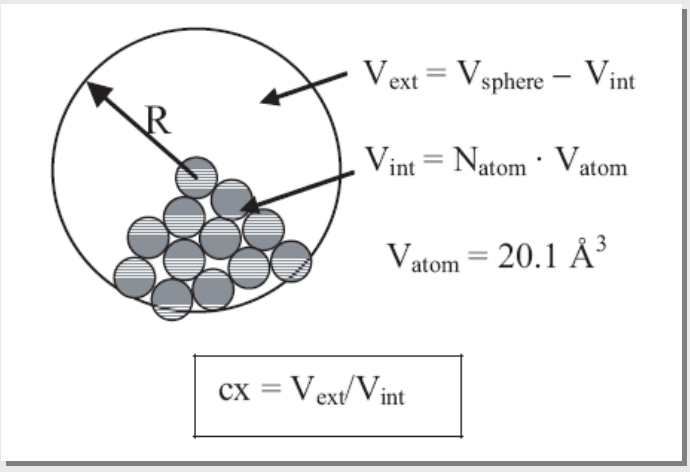
- Si existeixen altres àtoms que penetren dins l´esfera, el diàmetre d´aquesta es redueix fins que sigui tangent a l´àtom amb més penetrància, sempre i quan el radi de l´esfera no sigui inferior a 1Å. El resultat és una esfera de radi x (major a 1Å) per un parell d´àtoms (b)
- Es repeteixen els passos anteriors per tots els parells d`àtoms, d´aquesta manera s´obtenen un seguit d´esferes (c).
- Les esferes es situen en una graella tridimensional (d).
- El conjunt d´esferes es converteixen en polígons aproximats (e).
- El conjunt de polígons dóna una idea de la regió que forma la cavitat (f).

L´algorisme detallat es pot trobar a l´article original [21], i el programa pot ser descarregat de la pàgina web http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html.

CX

CX és una aplicació ràpida i senzilla que calcula l´índex de protrusió dels àtoms pesats d´una proteïna. L´algorisme és molt senzill, i es pot resumir en els següents passos:

Figura 2.5. Esquema del càlcul de volums interior i exteriors. Font: Bioinformatics. 2002 Jul;18(7):980-4.



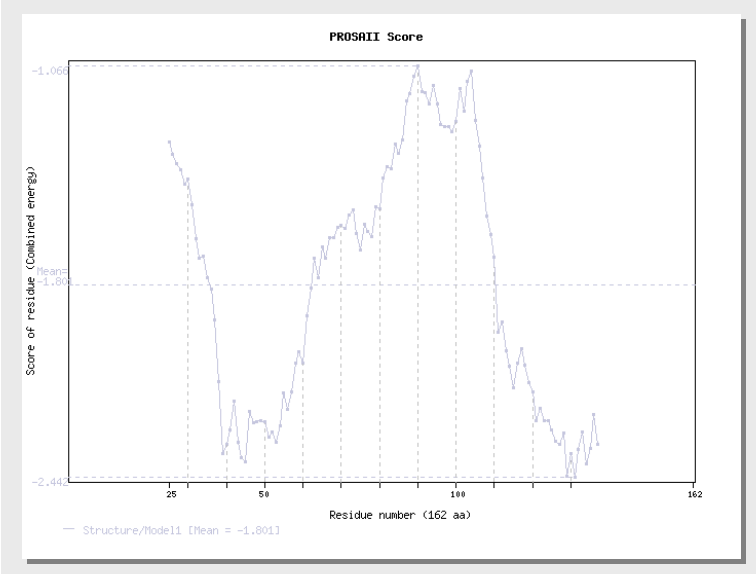
- A cada àtom pesat es fixa una esfera de radi concret.
- Seguidament es comptabilitza el volum ocupat per aquest àtom juntament amb els àtoms veïns que queden dins l´esfera. Aquest volum (V_{int}) es calcula multiplicant el número de residus que hi ha dins l´esfera pel volum mitjà d´un àtom pesat en una proteïna, que és $20.1 \pm 0.9 Å^3$ [22].
- L´índex de protrusió CX es representa com el quocient entre el volum lliure de l´esfera i l´ocupat pels àtoms.

Es poden veure més detalls sobre el programa a l´article original [23]. El programa pot ser descarregat lliurement de ftp://ftp.icgeb.trieste.it/pub/CX.

PROSA

Prosa (Protein Structure Analysis) és una aplicació que s´utilitza per valorar la qualitat de les estructures de proteïnes per mitjà de l'ús de potencials estadístics. Prosa assigna un potencial estadístic a cada residu de la proteïna a avaluar, en base a les distàncies que aquest presenta respecte la resta de residus. El perfil energètic local de la cadena permet veure quines regions de la proteïna contenen errors; aquestes parts correspondran a regions amb energies elevades (generalment superiors a zero).

Figura 2.6. Exemple de sortida del programa Prosa. Les zones positives corresponen a parts de la predicció errònies.



Els potencials estadístics de cada residu han estat calculats utilitzant proteïnes del PDB [24]. El detall del procediment es pot veure a l´article original [25].

Prosa ha estat àmpliament utilitzat tant en la validació d´estructures resoltes experimentalment, com en la de prediccions. El programa pot ser descarregat de http://www.came.sbg.ac.at/typo3/.

2.4. MODELLER

Modeller [26] és un programa de modelatge comparatiu dissenyat per trobar l´estructura tridimensional més probable per una seqüència donat un alineament de seqüència amb una estructura homòloga (mirar *Introducció*).

L´alineament de seqüències serveix de taula d´equivalències, de tal manera que Modeller pot prendre les restriccions espacials de l´estructura coneguda, i imposar-les a la seqüència a modelar. La seqüència resultant és sotmesa a una optimització molecular per tal d´evitar certes violacions geomètriques i energètiques.

Com en tots els mètodes de modelat comparatiu, la qualitat del model resultant depèn majoritàriament de l´alineament [27]; de tal forma que només les zones alineades tindran una estructura fiable; això fa que la resta de seqüència -que molt sovint correspon a *loop* o zones poc conservades- no puguin ser modelades de forma correcta. Modeller incorpora una funcionalitat limitada per predicció *ab initio* d´aquestes regions emprant tècniques de dinàmica molecular *in vacuo*. No obstant això cal remarcar que tot i aquestes funcionalitats, aquelles regions no alineades sempre seran menys fiables que la resta (mirar *Introducció*).

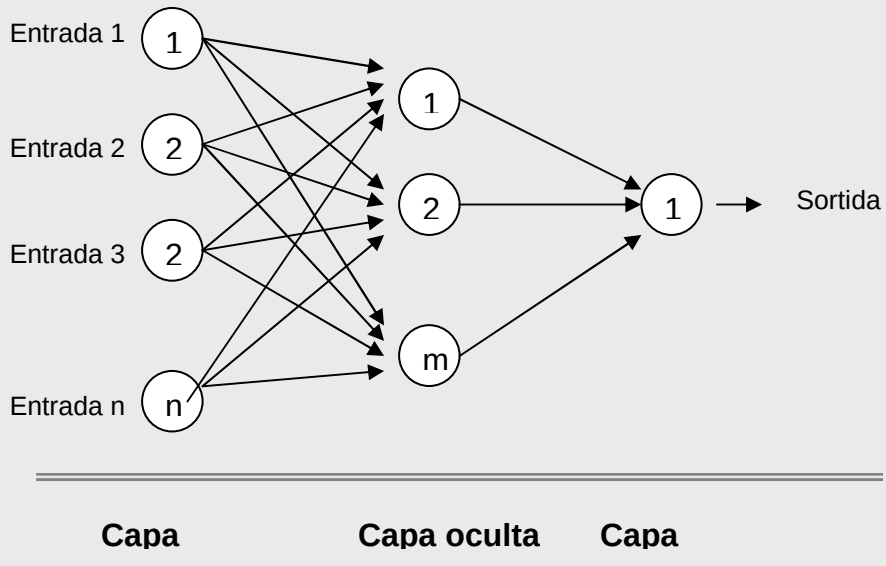
Es pot consultar el funcionament detallat del programa a l´article original [26]. Pel que fa al programa pot ser descarregat de forma gratuïta de la pàgina http://salilab.org/modeller.

2.5. XARXES NEURONALS

Les xarxes neuronals són eines d´intel.ligència artificial que s´apliquen habitualment als problemes de reconeixement de patrons, i que intenten simular les propietats i comportament dels sistemes neuronals biològics a través de models matemàtics [28]. L´objectiu és aconseguir que la xarxa neuronals doni respostes similars a les que donaria el cervell, però amb l´eficiència computacional dels ordinadors.

Una xarxa neuronal es composa d´unitats bàsiques o neurones. Cada neurona rep un seguit d´entrades a través d´interconnexions, i emet una sortida; tanmateix com una neurona biològica.

Figura 2.7. Esquema d´una xarxa neuronal amb *n* neurones d´entrada, i *m* neurones a la capa oculta. Cada cercle representa una neurona, i cada fletxa una interconnexió neuronal.



Segons el patró de connexions presents podem trobar xarxes neuronals:

- *Feed-forward*, on totes les senyals van des de la capa d´entrada fins la sortida, sense existir cicles ni connexions entre neurones d´una mateixa capa.
- *Recurrents*, presenten al menys un cicle tancat de connexió neuronal.

També segons el número de capes de neurones que existeixen podem parlar de xarxes neuronals *monocapa* o *multicapa*.

APRENENTATGE I FUNCIONAMENT DE LES XARXES NEURONALS

Seguint amb el símil amb les neurones biològiques, les xarxes neuronals per poder ser utilitzades han de ser prèviament entrenades. Si ens fixem per exemple en una xarxa neuronal utilitzada per classificar patrons en dues categories, l´aprenentatge optimitza els paràmetres del model matemàtic de la xarxa neuronal de tal manera que permeti diferenciar entre patrons diferents.

Els models d´aprenentatge més usuals són: aprenentatge supervisat i aprenentatge no-supervisat. En l´aprenentatge supervisat es necessita un conjunt de dades d´entrada (patrons) ja classificats. En l´aprenentatge no-supervisat no és necessari disposar d´aquest conjunt de dades.

VALIDACIÓ CREUADA

Per comprovar la fiabilitat del resultat de les xarxes neuronals se sol utilitzar la tècnica de la validació creuada [29]. En aquesta tècnica el conjunt de dades es divideix a l´atzar en dos subconjunts, entrenament i test. El primer subconjunt s´utilitza per entrenar la xarxa neuronal, i el segon per mesurar la capacitat de predicció adquirida.

PROBLEMA DEL DESBALANÇ DE CLASSES

Es pot donar el cas que vulguem classificar patrons en dues classes, utilitzant un mètode d´entrenament supervisat, on per la naturalesa de les dades el conjunt de dades d´entrenament estigui altament desbalancejat (molts més patrons d´una classe que de l´altra). Quan això succeeix el poder classificador de les xarxes neuronals decreix enormement, ja que tendeixen a classificar-ho tot com a patrons pertanyents a la classe majoritària.

Una solució al problema és fer un re-mostreig de les dades al conjunt d´entrenament, i equiparar les dades de la classe menys poblada a les de la classe superpoblada [30]. Existeixen dues alternatives:

- Remostratge negatiu: es treuen a l´atzar dades del conjunt superpoblat, fins aconseguir la proporció respecte la classe menys poblada desitjada.
- Remostratge positiu: es dupliquen a l´atzar dades del conjunt menys poblat, fins aconseguir la proporció respecte la classe superpoblada desitjada.

MESURA DE RENDIMENT DE LES XARXES NEURONALS

La mesura del rendiment de les xarxes neuronals es fa utilitzant magnituds habitualment utilitzades en sistemes de classificació o diagnòstic. Aquestes són exactitud, precisió, sensibilitat i especificitat.

$$\text{Exactitud} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precisió} = \frac{tp}{tp + fp}$$

$$\text{Sensibilitat} = \frac{tp}{tp + fn}$$

$$\text{Especificitat} = \frac{tn}{tn + fp}$$

On *tp* són les prediccions positives encertades; *tn* les prediccions negatives encertades; *fp* les prediccions positives que s´han fallat i *fn* les prediccions negatives que s´han fallat.

Una altra mesura que dóna una idea global de les anteriors és el coeficient de correlació de Matthews (mcc). Es calcula amb la següent fórmula:

$$\text{Mcc} = \frac{tp.tn - fp.fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$$

En ocasions pot interessar donar les mesures anteriors segons la fiabilitat de la sortida de la xarxa neuronal, per exemple el coeficient de Matthews possiblement serà millor si ens fixen en dades de fiabilitat alta, etc. Per fer-ho s´utilitza l´índex de fiabilitat [31]; aquest índex es calcula a partir del resultat de sortida de la xarxa neuronal segons la següent fórmula: integer(abs(NN_{output} – 0.5)x20), i va de 0 a 10.

XARXA NEURONAL UTILITZADA

En els diversos estudis s´ha utilitzat una xarxa neuronal tipus *feed-forward*, programada per Adrian Sheperd (University College of London, 1999). Els paràmetres d´execució es troben a la següent taula:

Taula 2.3. Paràmetres utilitzats en l´execució de la zarza neuronal

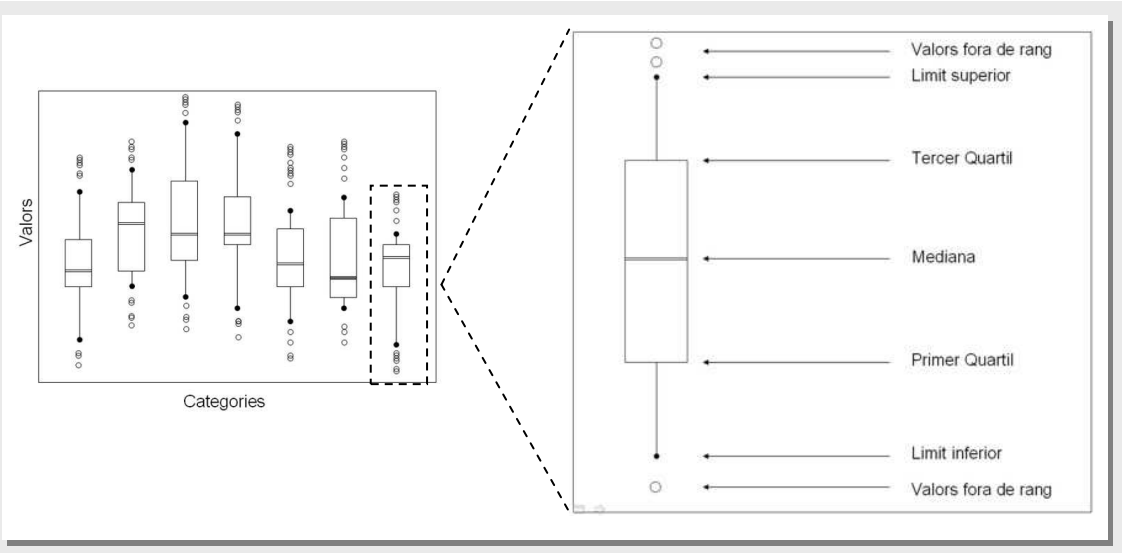
Paràmetre	Descripció	Valor
-H	Arquitectura	0 0 (perceptró lineal)
-A	Rang d´inicialització dels pesos	0.1
-S	Llavors per la inicialització dels pesos	1
-M	Mètode d´entrenament	Conjgrad (gradients conjugats escalars)
-g	Tolerància error global	0.01
-m	Nº màxim d´èpoques	500
-e	Llistar error cada n èpoques	50
-i	Llistar rendiment cada n èpoques	25
-w	Llistar pesos cada n èpoques	50

2.6. REPRESENTACIONS GRÀFIQUES

DIAGRAMES DE CAIXES O BOXPLOTS

Una forma eficaç de representar conjunts de dades es tenir en compte la distribució dels seus valors. Un exemple de representació que satisfà aquest criteri són els diagrames de caixes, o *boxplots*. Al llarg dels estudis que conformen aquesta tesi han estat àmpliament utilitzats, per això tot seguit s´expliquen en detall.

Figura 2.8. Exemple i informació d´una representació de caixes (boxplot).



Al l´esquema superior es pot veure les parts o informació que aporta un *boxplot*. La doble línea que es troba dins la caixa correspon a la mediana. El rectangle en si representa el rang interquartil (IQR), i va des del primer (percentil 25%) fins al tercer quartil (percentil 75%). Les línies que marquen els límits equivalen a:

- Límit superior: valor màxim – tercer quartil. Si aquest valor és superior a 1.5 vegades l´IQR, aleshores el límit superior equival a 1.5 vegades IQR
- Límit inferir: primer quartil – valor mínim. Si aquest valor és superior a 1.5 vegades l´IQR, aleshores el límit superior equival a 1.5 vegades IQR.

Els punts que queden per sobre o per sota dels límits són els valors fora de rang, també coneguts com a *outliers*.

VISUALITZACIÓ DE MOLÈCULES

Per la visualització de proteïnes s´han utilitzat fonamentalment dues aplicacions: VMD i Pymol. Ambdós són programes visuals que suporten gran varietat de plataformes, d´ús senzill i intuïtiu.

VMD

VMD, o Visual Molecular Dynamics, és un programa de visualització molecular desenvolupat per la Universitat d´Illinois, que permet mostrar, animar i analitzar sistemes biomoleculars de forma tridimensional.

Suporta sistemes operatius Windows, Linux i MacOS; i la seva distribució és gratuïta (com el codi font). Per més detall sobre les característiques, funcionament i descàrregues es pot consultar la pàgina web http://www.ks.uiuc.edu/Research/vmd/.

PYMOL

Pymol, al igual que VMD, és un altre programa de visualització molecular desenvolupat per DeLano Scientific.

Existeix per als sistemes operatius Windows, Linux i MacOS; i la seva distribució és també gratuïta. Es pot trobar informació detallada sobre característiques, funcionament i descàrrega a http://pymol.sourceforge.net.

2.7. REFERÈNCIES

1. Westbrook, J., et al., *The Protein Data Bank and structural genomics.* Nucleic Acids Res, 2003. **31**(1): p. 489-91.
2. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.
3. Orengo, C.A. and W.R. Taylor, *SSAP: sequential structure alignment program for protein structure comparison.* Methods Enzymol, 1996. **266**: p. 617-35.
4. Redfern, O.C., et al., *CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures.* PLoS Comput Biol, 2007. **3**(11): p. e232.
5. Swindells, M.B., *A procedure for detecting structural domains in proteins.* Protein Sci, 1995. **4**(1): p. 103-12.
6. Holm, L. and C. Sander, *Parser for protein folding units.* Proteins, 1994. **19**(3): p. 256-68.
7. Siddiqui, A.S. and G.J. Barton, *Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions.* Protein Sci, 1995. **4**(5): p. 872-84.
8. Levitt, M. and C. Chothia, *Structural patterns in globular proteins.* Nature, 1976. **261**(5561): p. 552-8.
9. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.
10. Hughey, R. and A. Krogh, *Hidden Markov models for sequence analysis: extension and analysis of the basic method.* Comput Appl Biosci, 1996. **12**(2): p. 95-107.
11. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
12. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Cryst, 1976. **32**: p. 922-923.

13. Carugo, O. and S. Pongor, *A normalized root-mean-square distance for comparing protein three-dimensional structures.* Protein Sci, 2001. **10**(7): p. 1470-3.

14. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.

15. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.

16. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.

17. Alexandrov, V., *LSQRMS.* Yale, 2000.

18. Alexandrov, N.N. and N. Go, *Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins.* Protein Sci, 1994. **3**(6): p. 866-75.

19. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.

20. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

21. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. **13**(5): p. 323-30, 307-8.

22. Richards, F.M., *The interpretation of protein structures: total volume, group volume distributions and packing density.* J Mol Biol, 1974. **82**(1): p. 1-14.

23. Pintar, A., O. Carugo, and S. Pongor, *CX, an algorithm that identifies protruding atoms in proteins.* Bioinformatics, 2002. **18**(7): p. 980-4.

24. Bernstein, F.C., et al., *The Protein Data Bank. A computer-based archival file for macromolecular structures.* Eur J Biochem, 1977. **80**(2): p. 319-24.

25. Sippl, M.J., *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.* J Mol Biol, 1990. **213**(4): p. 859-83.

26. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.

27. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.

28. Basheer, I.A. and M. Hajmeer, *Artificial neural networks: fundamentals, computing, design, and application.* J Microbiol Methods, 2000. **43**(1): p. 3-31.

29. Krishnan, V.G. and D.R. Westhead, *A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function.* Bioinformatics, 2003. **19**(17): p. 2199-209.

30. Japkowicz, N. and S. Stephen, *The class imbalance problem: a systematic study.* Intelligent Data Analysis Journal, 2002. **6**.

31. Shepherd, A.J., D. Gorse, and J.M. Thornton, *Prediction of the location and type of beta-turns in proteins using neural networks.* Protein Sci, 1999. **8**(5): p. 1045-55.

CAPÍTOL 3.

IDENTIFICACIÓ DE LA FAMÍLIA ESTRUCTURAL DE PREDICCIONS DE NOVO

3.1. INTRODUCCIÓ

Tal i com s´ha tractat a la introducció, la predicció de l´estructura tridimensional de les proteïnes continua essent un dels reptes de la biologia molecular actual, i tot i que hi ha mètodes com el modelat per homologia que donen bons resultats, la predicció estructural a partir de la seqüència és encara quelcom per resoldre. Els principals motius són que l´espai conformacional a explorar és molt gran [1], i que ara per ara no disposem de funcions d´energia 100% eficients que permetin dir si una estructura predita és correcta o no.

No obstant això, els progressos en aquest camp són importants. David Baker i col.laboradors per exemple, utilitzant Rosetta@home, que és xarxa computacional distribuïda basada en el *Berkeley Open Infrastructure Network Computer protocol* (http://boinc.bakerlab.org/rosetta/), han estat capaços d´obtenir diverses prediccions estructurals *de novo* a resolucions properes als 2Å [2]; el problema és que el procediment és computacionalment molt costós, i únicament és aplicable a proteïnes de menys de 100 residus. Si ens centrem en situacions computacionals més convencionals, el normal és obtenir models dins un rang de resolucions molt més ampli. Així doncs, la realitat avui en dia és que tot i que és possible generar prediccions força correctes, encara estem lluny d´aconseguir prediccions bones de forma general, tal i com ho corroboren els resultats dels darrers experiments CASP i ha estat comentat a la introducció.

El problema que apareix en aquest cas és que la utilitat d´aquestes prediccions depèn directament de la qualitat; així per exemple una predicció de 6-8Å possiblement sigui útil per estudiar l´efecte de mutacions sobre el plegament, però en la majoria dels casos serà de poca utilitat en el disseny de fàrmacs. Una solució, tal i com es debatrà més profundament en aquest i els dos capítols posteriors, és sotmetre aquestes prediccions de baixa resolució a tècniques de refinament estructural [3-5].

Al llarg d´aquests tres primers capítols d´aquesta tesi presentem una aproximació senzilla al que creiem que seria un protocol de refinament lògic. Aquest protocol es podria resumir en les següents etapes:

1. Determinació del a família estructural d´un conjunt de prediccions *de novo* de baixa resolució.
2. Determinació de la qualitat local dels models
3. Utilització de la informació estructural de família per millorar les prediccions inicials.

Al llarg del capítol es discutirà en detall el primer pas; més concretament es tractarà un mètode de determinació de la família estructural basat en l'ús d´eines de comparació estructural i xarxes neuronals. Aquest treball es basa en resultats prèviament obtinguts en el grup [6].

Pel que fa al segon i tercer punts, seran tractats amb més detall al llarg del capítol 4 i especialment al capítol 5.

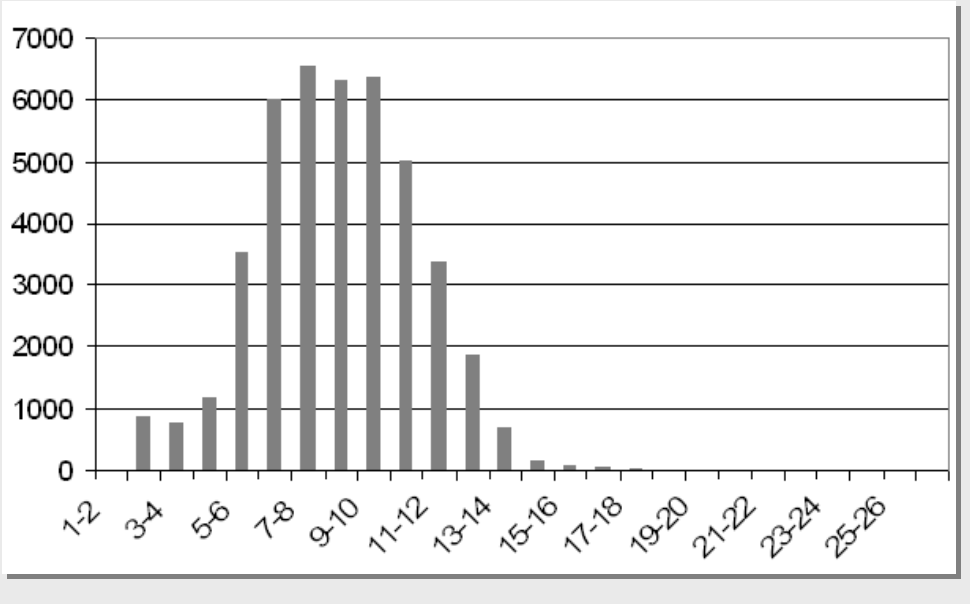
3.2. METODOLOGIA

CONJUNT DE PREDICCIONS *DE NOVO*

S´ha utilitzat un conjunt de prediccions *de novo* realitzades amb el programa Rosetta [7]. En total disposem de 92 proteïnes amb 999 prediccions cadascuna. D´aquestes només 84 presentaven codi CATH en el moment que es va fer l´estudi. És indispensable que presentin codi CATH, ja que és el que ens serveix per establir la seva topologia.

En la figura 3.1 es pot apreciar el perfil de RMSDs de les 999 prediccions per a 43 d´aquestes proteïnes.

Figura 3.1. Distribució d´RMSD per als 999 models de 43 de les proteïnes estudiades. La distribució és extrapolable a les 92 proteïnes.



DOMINIS SREP DE CATH

Com a base de dades de dominis estructurals s´ha utilitzat la llista de representants S (Sreps) de la base de dades CATH versió 2.51. Aquesta llista consta de 4023 dominis.

La topologia dels dominis es va assignar a partir del seu codi CATH (nivell T). Per més detall sobre els nivells CATH mirar el capítol de *Metodologia general*.

EINES DE COMPARACIÓ ESTRUCTURAL

S´han utilitzat cinc programes de comparació estructural típicament utilitzats en l´entorn de predicció i anàlisi estructural: SSAP [8], LGA [9], MAMMOTH [10], CE [11] i LSQ [12]. El detall de com funcionen es pot consultar al capítol de *Metodologia general*.

XARXA NEURONAL

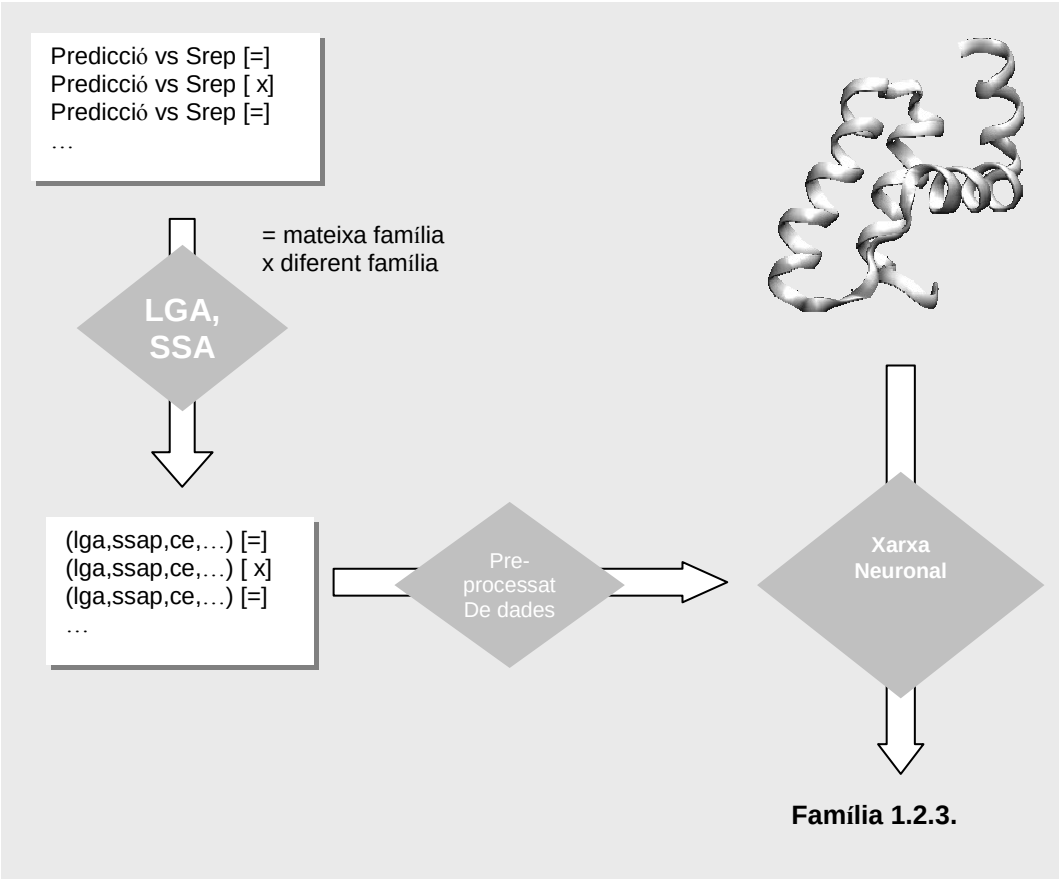
S´ha utilitzat una xarxa neuronal tipus *feed-forward*, programada per Adrian Sheperd (University College of London, 1999). Els paràmetres d´execució es poden consultar al capítol de *Metodologia general*.

3.3. RESULTATS

PROTOCOL D´IDENTIFICACIÓ DE FAMÍLIA ESTRUCTURAL

La idea del protocol d´identificació és senzilla: comparem estructuralment una predicció *de novo* amb un conjunt de dominis CATH; obtenint un vector de similituds. Seguidament s´entrena una xarxa neuronal de tal manera que sigui capaç de relacionar aquest vector de similituds entre predicció i domini estructural amb dues situacions: la predicció i el domini estructural pertanyen/no pertanyen a la mateixa família estructural. D´aquesta manera si la xarxa té un bon rendiment, comparant una predicció *de novo* amb un conjunt de dominis estructurals podríem ser capaços d´identificar la seva família estructural. El protocol es pot veure esquematitzat en la figura 3.2:

Figura 3.2. Esquema general del protocol d´identificació de família estructural. Consta de tres passos: comparació estructural, preprocessat de dades i ús de la xarxa neuronal.



Tot seguit es passen a descriure els tres passos principals:

Comparacions estructurals.

En el primer pas s´efectuen les comparacions estructurals de cadascuna de les prediccions *de novo* amb cadascun dels representants S de CATH (4023 dominis), utilitzant els diversos mètodes de comparació estructural: MAMMOTH, LGA, LSQ, CE i SSAP. Així doncs, per cada proteïna tindrem 999x4023 comparacions. Per cadascuna d´aquestes comparacions podem expressar les cinc puntuacions com un vector de similituds, que posteriorment s´utilitzarà per entrenar la xarxa neuronal.

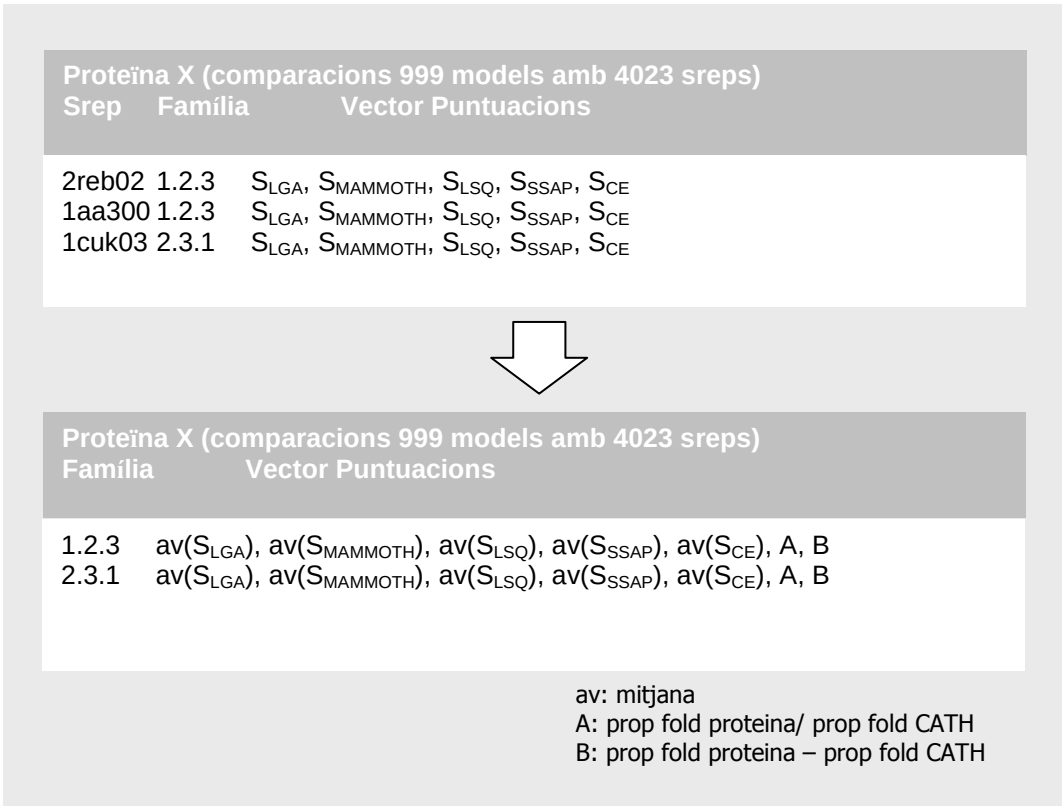
Pre-processat de les dades.

Les dades de puntuacions derivades del pas anterior no s´utilitzen directament per entrenar les xarxes neuronals, sinó que es sotmeten a passos de preprocessat:

- Filtre per mida de l´alineament: només es conserven aquelles comparacions on hi ha més d´un 60% de residus alineats per a tots els mètodes de comparació. El motiu d´aplicar aquest filtre és eliminar aquelles comparacions entre prediccions *de novo* i dominis massa llunyanes.

- Agrupació de vectors per família: per tal de reduir la complexitat els vectors de similituds de les prediccions de cada proteïna s´han agrupat segons la seva família estructural, entenent per família estructural el nivell T de CATH. A més s´han afegit dos paràmetres més que permeten comptabilitzar la proporció de la família estructural a CATH i al conjunt de proteïnes d´estudi. En la següent figura es pot veure resumit:

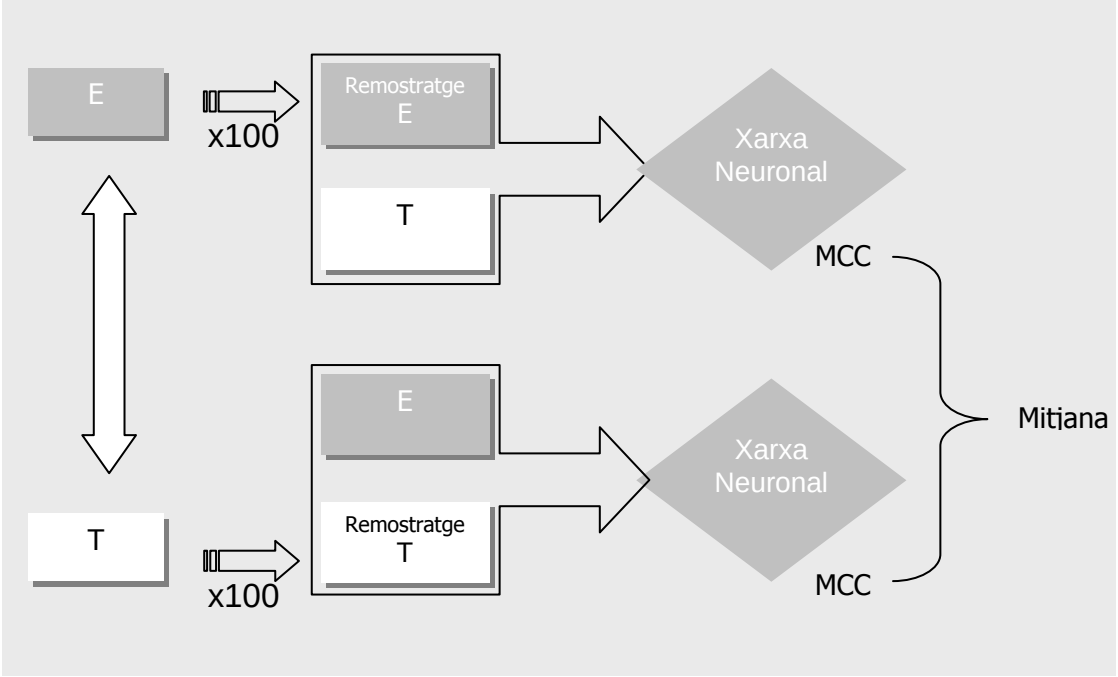
Figura 3.3. Agrupació dels vectors de similitud per família estructural. Per totes les comparacions que impliquen un domini de la mateixa família estructural es fan les mitjanes aritmètiques (av). S´afegeixen dos paràmetres A i B que aporten informació sobre la proporció de la família estructural en qüestió al conjunt de proteïnes i a la base de dades CATH.



Ús de la xarxa neuronal.

Es va seguir un protocol de validació creuada per evitar, dins del possible, una sobreestimació de la capacitat predictiva del mètode. Aquest protocol fou acompanyat d´un remostratge de les dades per tal de corregir efectes de tipus composicional (mirar *Metodologia general*). El següent esquema resumeix aquest procés:

Figura 3.4. Esquema de protocol seguit per utilitzar la xarxa neuronal: disposem d´un conjunt d´entrenament (E) i un conjunt de prova (T). En primer lloc es remostreja el conjunt E, i es prova la xarxa neuronal sobre el conjunt T; el rendiment de la xarxa es mesura amb el MCC. Seguidament es remostreja el conjunt T i es prova la xarxa neuronal sobre el conjunt E; el rendiment de la xarxa novament es mesura amb el MCC. S´amitjanen ambdós coeficients. El cicle es repeteix 100 cops. El MCC final correspon al promig individual de cada execució.



Disposem d´un conjunt d´entrenament i de prova, compostos per 38 i 37 proteïnes respectivament.

- Sotmetem els conjunts d´entrenament i prova a remostratge positiu/negatiu, amb proporció de classe minoritària:majoritària 1:2.
- Entrenem la xarxa neuronal sobre el conjunt remostrejat, i provem l´eficàcia sobre l´altre. Cal notar que apliquem validació creuada, per tant en un cas entrenem sobre el grup d´entrenament remostrejat i fem la predicció sobre el grup de prova, i en l´altre entrenem sobre el grup de prova remostrejat i fem la predicció sobre el grup d´entrenament.
- Es pren una primera mitjana aritmètica dels paràmetres de rendiment de la xarxa (veure *Metodologia general).*

- Es repeteixen els passos anteriors 100 cops, i es pren una segona mitjana aritmètica com a mesura global de la xarxa.

CONTRAST DEL PROTOCOL

En la següent taula es mostren les dades de rendiment de les xarxes neuronals:

Taula 3.1. Rendiment de la xarxa neuronal utilitzada.

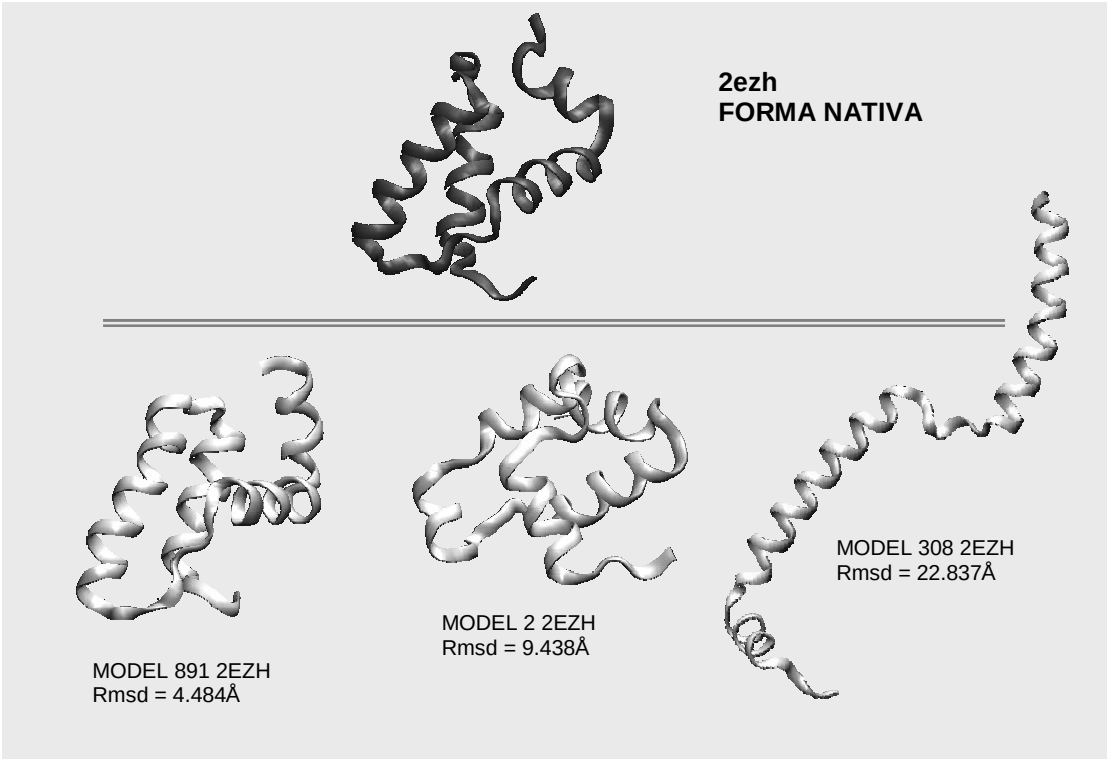
	Mesura	Remostreig Negatiu	Remostreig Positiu
Rendiment global	MCC	0.34±0.05	0.41±0.02
	MCC fiabilitat < 5	0.17±0.08	0.23±0.06
	MCC fiabilitat > 5	0.51±0.07	0.57±0.03
Predicció Diferent Família	Sensibilitat	92%	95%
	Precisió	98%	98%
	Millora sobre l´atzar	19%	26%
Predicció Mateixa Família	Sensibilitat	63%	63%
	Precisió	21%	28%
	Millora sobre l´atzar	60%	61%

Tenint en compte aquests resultats podem dir que els resultats són moderadament positius, tot i que resulta evident que el protocol proposat encara requereix la introducció de millores. A continuació es senyalen els aspectes més destacables dels resultats:

- El mètode és més fiable dient quan una predicció NO pertany a una determinada família estructural.
- Tot i que la sensibilitat, precisió o millora sobre l´atzar en alguns casos tenen valors elevats, si ens fixem en els coeficients de Matthews veiem que són relativament reduïts; això demostra que tot i que el protocol proposat s´aproxima a una solució del problema encara es troba lluny de resoldre´l.
- Les tècniques de remostratge no són suficients per millorar l´eficiència de la xarxa neuronal.

De totes maneres hi ha casos particulars on el mètode funciona i permet determinar la família estructural d´un conjunt de prediccions. Un exemple és 2ezh: la xarxa neuronal és capaç de reconèixer la família estructural en els 100 cicles (veure l´apartat de *Ús de la xarxa neuronal*). A la figura 3.5 es pot apreciar la forma nativa i tres prediccions *de novo* de 2ezh.

Figura 3.5. Exemple de proteïna per la qual la xarxa neuronal és capaç de determinar-ne la família estructural. S´ha considerat que la xarxa neuronal funciona quan en els 100 cicles és capaç de dir-nos que les prediccions de 2ezh pertanyen a la família estructural 1.10.10.



3.4. DISCUSSIÓ

En aquest capítol s´ha presentat un mètode per a la identificació de la família estructural d´un conjunt de prediccions *de novo*, basat en mètodes de comparació estructural i xarxes neuronals.

Els resultats obtinguts mostren que tot i que és una aproximació al problema, el rendiment és encara moderat, i no aporta una solució definitiva al problema. Seguidament s´exposen alguns dels motius que creiem que poden tenir pes sobre el funcionament del protocol, i possibles solucions que podrien ser aplicades en un futur.

En primer lloc caldria parlar del descriptor. Podria ser que el vector de 7 valors utilitzat com a descriptor (veure *Resultats*) no sigui suficient per poder identificar la família estructural d´una predicció. Una solució seria investigar sobre l'ús d´altres variables com per exemple identitat de seqüència, presència d´estructura secundària, accessibilitat,... El problema que això comporta és que a mesura que incrementem la mida i complexitat del descriptor es fa necessari un volum de dades molt major sobre el que treballar (ex. Un conjunt de prediccions més ampli, per més de 85 proteïnes). Per altra banda, millorar el descriptor en base als resultats que varem obtenir pot conduir a un cas de sobre ajustament de la xarxa al conjunt de dades utilitzades, amb el que donaria molt bons resultats per aquestes, però amb dades independents els resultats podrien empitjorar substancialment.

En segon lloc tenim la diversitat de la qualitat estructural del conjunt de prediccions. A la figura 3.1 podem veure com la majoria de les prediccions utilitzades estan per sobre dels 8Å. Es podria donar el cas que una predicció tingués una similitud baixa amb una proteïna de la mateixa família estructural, i en canvi tingués una similitud alta amb una proteïna d´una altra família. Aquest fet tindria un efecte negatiu sobre la validesa del descriptor utilitzat, ja que ens trobaríem per exemple vectors de similitud alts per

situacions on una predicció i un domini comparteixen la mateixa família estructural, però també en el cas contrari. La solució a aquest problema és complicada, i està principalment associada a la millora dels mètodes de predicció *de novo*, fins al punt de ser capaços de generar prediccions amb plegament natiu, o proper al natiu.

En tercer i últim lloc, hom podria pensar en utilitzar altres sistemes d´intel.ligència artificial. En certs estudis (buscar referència) s´ha comparat el rendiment de les xarxes neuronals amb altres algorismes d´intel.ligència artificial com poden ser els Support Vector Machines, o (buscar altre exemple). En col.laboració amb Oliver Redfern (University College of London) hem realitzat proves utilitzant els primers, però els resultats no han estat superiors. Aquestes proves, junt amb proves d´execució de les xarxes neuronals amb altres arquitectures (dades no mostrades) corroboren que la complexitat del problema no rau en l´algorisme de reconeixement utilitzat, sinó en els punts tractats anteriorment.

Resumint, el problema de la identificació de la família estructural continua essent un tema obert, i fins que no es posi solució a alguns dels punts comentats anteriorment serà difícil resoldre el problema, si més no des d´una perspectiva com la que es pretén en aquest estudi.

3.5. REFERÈNCIES

1. Levinthal, C., *ARE THERE PATHWAYS FOR PROTEIN FOLDING?* Journal de Chimie Physique, 1968. **65**(1): p. 2.
2. Das, R., et al., *Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home.* Proteins, 2007. **69 Suppl 8**: p. 118-28.
3. Eyrich, V.A., D.M. Standley, and R.A. Friesner, *Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set.* J Mol Biol, 1999. **288**(4): p. 725-42.
4. Monge, A., et al., *Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models.* J Mol Biol, 1995. **247**(5): p. 995-1012.
5. Ortiz, A.R., A. Kolinski, and J. Skolnick, *Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments.* J Mol Biol, 1998. **277**(2): p. 419-48.
6. de la Cruz, X., I. Sillitoe, and C. Orengo, *Use of structure comparison methods for the refinement of protein structure predictions. I. Identifying the structural family of a protein from low-resolution models.* Proteins, 2002. **46**(1): p. 72-84.
7. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.
8. Orengo, C.A. and W.R. Taylor, *SSAP: sequential structure alignment program for protein structure comparison.* Methods Enzymol, 1996. **266**: p. 617-35.
9. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.
10. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.

11. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.

12. Alexandrov, V., *LSQRMS.* Yale, 2000.

CAPÍTOL 4.

IDENTIFICACIÓ DE ZONES CORRECTES EN PREDICCIONS *DE NOVO*

4.1. INTRODUCCIÓ

El ràpid progrés que ha experimentat el camp de la predicció estructural *de novo* en els darrers anys ha impulsat el desenvolupament en paral.lel, dels mètodes d´avaluació de la qualitat de les prediccions [1, 2]. Les raons es poden resumir en els següents punts:

- La qualitat d´una predicció determinarà la seva aplicació [3]; per exemple una predicció de qualitat mitjana podrà ser emprada per l'estudi de l´efecte de mutacions, però en la majoria dels casos no podrà ser emprada en camps com el disseny de fàrmacs (mirar *Capítol 7*).
- Les tecnologies de predicció *de novo* cada cop són més accessibles, i el perfil de l´usuari ha canviat, de manera que cada cop més gent, d´àmbits més o menys allunyats de la bioinformàtica estructural, pot utilitzar algun programa de predicció estructural. Així doncs és important que disposin de mètodes per determinar la qualitat dels models obtinguts.

Històricament, els mètodes d´assignació de qualitat estructural varen ser desenvolupats per avaluar les estructures experimentals obtingudes per raig X o RMN. Tot i que existeixen diverses metodologies, genèricament podem parlar de dos tipus, segons l´aspecte de l´estructura de la proteïna analitzat [4, 5]:

- Estructura covalent (enllaços, angles d´enllaç i torsionals): programes com Procheck [6], What-Check [7].
- Estructura no covalent (patró de contactes no covalents i exposició al solvent): programes com Prosa [8], Anolea [9], Verify3D [10], etc.

Com que els paràmetres que defineixen la qualitat d´una estructura experimental són extrapolables a les prediccions estructurals, els mètodes anteriors han passat a emparar-se en el camp de la determinació de qualitat de prediccions. Així programes de modelat per homologia com Modeller recomanen com a part del seu protocol l'avaluació de models emprant Procheck o Prosa entre altres. Programes com Whatif [11] i Prosa [8] també són habitualment emprats en experiments CASP, juntament amb altres mètodes desenvolupats específicament per l´avaluació de prediccions estructurals, per exemple PCons [12, 13] entre altres programes desenvolupats per Arne Elofsson [14]. A part del camp de la determinació de qualitat purament dita, aquests programes han entrat en altres àmbits d´investigació, com per exemple identificació de la família estructural d´un seguit de prediccions [15, 16] tal i com s´ha vist en el capítol anterior.

En aquest capítol es descriu un mètode senzill per l´avaluació de les prediccions *de novo*, que denominarem SCLQA (**S**tructure **C**omparison-based **L**ocal **Q**uality **A**ssessment, o Assignació de Qualitat Local basada en Comparació Estructural). Es basa en una idea senzilla: comprovar si quan comparem una predicció amb un homòleg, els residus alineats tenen millor qualitat que la resta (figura 4.2). Aquesta idea està inspirada en:

- L´extens ús dels mètodes de comparació estructural en els experiments CASP: aquests mètodes són àmpliament emprats en aquests experiments per tal de determinar si les prediccions presentades pels participants tenen plegaments similars al natiu.
- Treballs previs on aquests mètodes de comparació estructural han estat aplicats en el camp de la predicció *de novo*; entre els quals trobem el comentat en el capítol 3. A part, altres treballs previs al grup [16] i altres laboratoris [15] han emprat aquests mètodes en la determinació de la família estructural d´un conjunt de prediccions.

Per tal d´extrapolar la validesa d´aquesta hipòtesi s´ha fet ús d´un total de 17180 prediccions *de novo* corresponents a 68 proteïnes, generades amb el programa Rosetta [17]. Tot i que existeixen altres programes capaços de generar prediccions amb bons resultats [1, 2] hem seleccionat Rosetta per que és un programa que ha obtingut les millors puntuacions en els darrers experiments CASP [1, 2, 18-20]; a més, un gran nombre de prediccions realitzades amb aquest programa es troben disponibles al servidor del grup de David Baker (http://robetta.bakerlab.org).

4.2. METODOLOGIA

PREDICCIONS *DE NOVO*

Les prediccions *de novo* han estat descarregades del servidor del laboratori de David Baker (http://depts.washington.edu/bakerpg/). El conjunt consta de 999 prediccions *de novo* per un total de 85 proteïnes.

El protocol SCLQA emprat requereix dues condicions:

- Ha d´existir un homòleg amb estructura: per satisfer aquesta condició hem imposat que les proteïnes havien de presentar algun homòleg que compartís la mateixa topologia segons CATH (nivell T).
- Les prediccions han de presentar el *fold* correcte: per satisfer la segona condició s´han exclòs tots aquells models que no presentaven el mateix *fold* que els homòlegs. A nivell tècnic, hem mantingut només aquelles prediccions amb una puntuació segons Mammoth superior a 5.25 quan es comparen amb un homòleg (aquest valor ha estat tret de proves realitzades amb aquest programa, i han estat emprades també per altres autors [15]).

El segon punt és computacionalment costós, ja que implica comparar cadascuna de les 999 prediccions per les 85 proteïnes amb els seus respectius homòlegs; és per aquest motiu que només s´ha dut a terme amb el programa Mammoth [21], que presenta una velocitat d´execució molt alta. Després d´aplicar aquests dos filtres el número total de prediccions disponibles baixa a 17180, pertanyents a 68 proteïnes (cal notar que no totes les proteïnes contribueixen amb el mateix número de models).

REPRESENTANTS CATH

Com s´ha comentat, el protocol SCLQA es basa en alinear un conjunt de models d´una proteïna amb els seus respectius homòlegs. La relació d´homologia ha estat assignada

emprant la base de dades de dominis CATH [22]. Pel propòsit del nostre estudi hem considerat que dues proteïnes són considerades homòlogues quan pertanyen al mateix grup T, o el que és el mateix, presenten la mateixa topologia. Així doncs, per una determinada proteïna, la llista d´homòlegs estarà confeccionada per tots aquells dominis que comparteixin els tres primers números del seu codi CATH (veure *Metodologia general*). Ja que emprar tots els dominis existents amb la mateixa Topologia implicaria utilitzar un conjunt massa extens d´homòlegs, s´han utilitzat només els representants de classe S. Aquests representants són aquells dominis que dins un nivell S presenten una major qualitat.

A la taula 4.1 es pot veure el conjunt d´homòlegs (representants S) que pertanyen a les diverses proteïnes emprades en l´estudi.

Taula 4.1. Proteïnes utilitzades en l´estudi, junt amb el número d´homòlegs corresponents.

Targets	#Homologs	Targets	#Homologs	Targets	#Homologs
1aa2	13	1hsn	8	1tit	211
1aa3	2	1jvr	1	1tul	1
1acf	21	1ksr	179	1utg	1
1ag2	3	1kte	61	1uxd	14
1ail	56	1leb	184	1vls	53
1aj3	37	1lfb	165	1who	253
1ark	71	1lis	2	1wiu	315
1ayj	14	1lz1	4	2acy	170
1bdo	23	1mbd	27	2ezh	181
1bor	3	1mzm	3	2ezk	171
1c5a	2	1nkl	3	2fdn	77
1cc5	28	1nre	1	2fha	13
1csp	87	1nxb	3	2fow	151
1ctf	6	1orc	1	2gdm	27
1ddf	14	1pal	60	2hp8	1
1eca	27	1pdo	735	2ncm	281
1erv	64	1pgx	64	2pac	22
1fbr	12	1pou	15	2ptl	68
1fwp	105	1qyp	5	2sn3	8
1gb1	65	1r69	15	4fgf	8
1gpt	2	1ris	157	5icb	62
1gvp	44	1sro	104	5pti	5
1hlb	27	1svq	11		

MÈTODES DE COMPARACIÓ ESTRUCTURAL

S´han escollit tres mètodes de comparació estructural d´ús comú: Mammoth [21], Ssap [23], i LGA [24]. Tots tres programes es troben als corresponents servidors, i es poden executar de forma simple. En tots tres casos els programes s´han executat emprant paràmetres per defecte.

Mammoth: ha presentat bons resultats puntuant prediccions *de novo* en diversos experiments CASP, i el temps d´execució és extremadament ràpid. Pensant sobretot en aquest darrer punt, Mammoth ha estat el mètode emprat per seleccionar aquelles prediccions que presentaven el mateix *fold* que el corresponent homòleg (veure apartat anterior). El programa genera una sortida on es mostra una puntuació per la comparació i una llista de residus alineats; per el nostre treball hem considerat com a residus alineats aquells que es trobaven a menys de 4Å.

Ssap: és el nucli de la base de dades CATH [22], ja que es fa servir per classificar els dominis en els diferents nivells jeràrquics que constitueixen la base de dades. De forma similar a Mammoth la sortida del programa és una puntuació i una llista de residus alineats; per al nostre estudi hem seleccionat aquells residus alineats que presentaven una puntuació superior a 10.

LGA: àmpliament emprat en experiments CASP. Com els dos anteriors dóna una puntuació i una correspondència entre residus alineats de les proteïnes comparades; hem considerat com a residus ben alineats aquells que estan per sota de 5Å.

Tots tres mètodes ens han donat resultats similars, és per això que per millorar la claredat només en alguns casos es mostraran els resultats per tots tres.

RMSD

La qualitat dels residus alineats s´ha determinat utilitzant l´RMSD com a variable descriptora [25]. Concretament, per cada mètode, després de comparar una predicció amb un homòleg s´ha extret el conjunt de residus alineats, i s´ha calculat l´RMSD sobre aquests. S´ha realitzat també el mateix càlcul sobre el total de residus, així tenim dues mesures que permeten comparar la qualitat de la part alineada, i la qualitat global de la predicció.

GDT_TS

GDT_TS és la mitjana dels valors GDT (Global Distance Test) calculats sobre 4 llindars, 1Å, 2Å, 4Å i 8Å. És una mesura emprada típicament per assignar la qualitat a prediccions *de novo*. GDT_TS ha estat calculat usant el procediment iteratiu descrit per Zemla [24]. Es troba explicat a l´apartat de Metodologia general. Permet detectar la presència de subestructures conservades.

MD, SLS i SAS

Per tal d´estudiar com les parts alineades o STR (Selected Target´s Residues) es distribueixen al llarg de la seqüència hem emprat tres paràmetres:

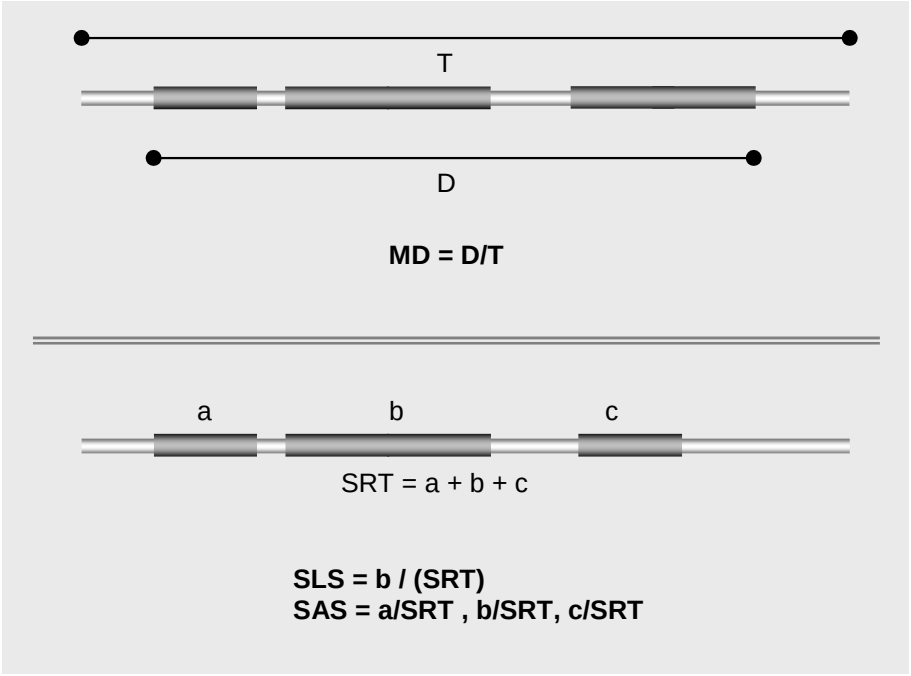
- MD o distància màxima: correspon a la longitud entre l´inici del primer fragment alineat i el final del darrer, normalitzada per la mida total de la seqüència. Dóna una idea sobre si l´STR es troba centrat a una zona de la proteïna, o pel contrari es distribueix per tota la seqüència. En aquest darrer cas MD tendirà a la unitat.

- SLS o mida del fragment alineat major: tenint en compte aquest paràmetre podem veure si els fragments dels STR són prou grans com per incloure

elements d´estructura secundària com hèlix o fulles beta. El valor de SLS està normalitzat per la mida de l´STR.

- SAS o distribució dels fragments de l´STR per mida: distribució de les mides dels fragments que conformen l´STR normalitzades per la mida de l´STR. Junt amb el paràmetre anterior donen una idea de la cobertura de l´STR respecte la seqüència de la proteïna (ATR).

Figura 4.1. Descriptors de la distribució dels STR al llarg de la seqüència: MD, SLS i SAS.



ÀREA DE SUPERFÍCIE ACCESSIBLE I ESTRUCTURA SECUNDÀRIA

La superfície accessible ha estat calculada amb el programa NACCESS [26], utilitzant una sonda de radi 1.4Å. S´ha considerat que un residu es troba accessible o enterrat si la seva accessibilitat relativa és inferior o superior a 20 % respectivament.

Pel que fa a l´estructura secundària dels residus, ha estat calculada amb el programa DSSP [27]. Aquest programa dóna per cada residu una predicció sobre el motiu estructural que presenta: hèlix, fulla beta, etc. Com es detallarà més endavant hem considerat dos categories de residus: els que pertanyen a una estructura periòdica (hèlix o fulla beta) i els que no.

IDENTITAT DE SEQÜÈNCIA

L´alineamet de seqüència ha estat calculat emprant l´algorisme estàndard de programació dinàmica desenvolupat per Needleman i Wunsch [28]. Pel que fa a la identitat, s´ha calculat dividint el número de residus alineats del mateix tipus pel número de residus alineats totals.

PROSA

Per tal d´obtenir una referència per als nostres valors s´ha fet ús del programa Prosa [8]. Actualment existeixen nombrosos programes que permeten determinar si una predicció és correcta; resultats de recents proves realitzades pel grup d´Arne Elofsson, junt amb resultats del darrer experiment CASP, mostren que el mètode Pcons [12, 13] és probablement la millor alternativa; no obstant això per avaluar la qualitat d´una predicció necessita un conjunt de models, i en aquest treball ens hem centrat en l´estudi de prediccions individuals. Per casos com el que es tracta en aquest capítol Prosa i Verify3D donen resultats correctes i força similars; simplement ens hem decantat per utilitzar Prosa ja que ha estat i continua essent un estàndard.

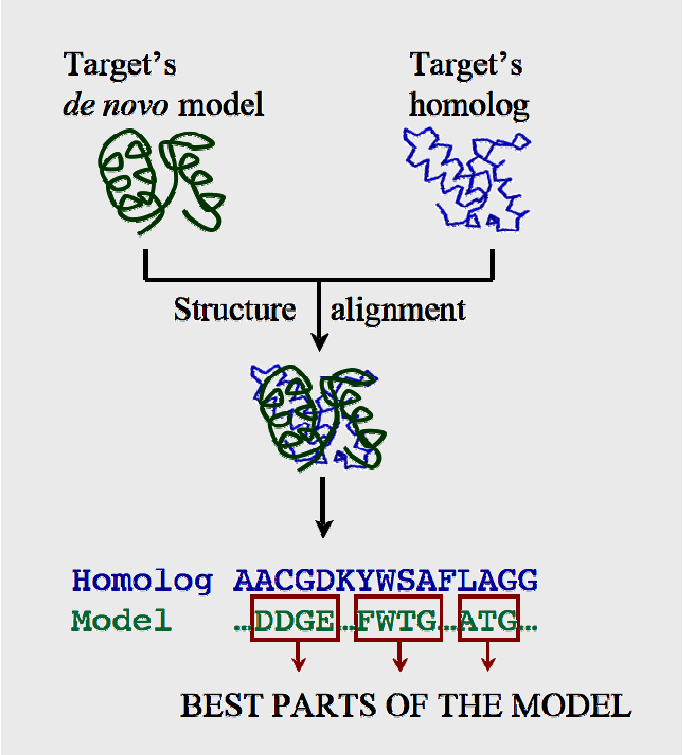
Es va utilitzar Prosa amb l´objectiu de trobar per cada predicció un conjunt de residus ben predits, així doncs, per cada predicció *de novo* es va efectuar un càlcul de potencials Prosa, de tal manera que els residus amb energia inferior a 0 es van considerar com a correctes (equivalents a residus ben alineats en el cas dels mètodes de comparació estructural). Aquests residus es van sotmetre de la mateixa manera a càlculs d´RMSD i GDT_TS. Es van fer proves baixant el llindar energètic a -1; no obstant, tot i que els resultats eren similars, el conjunt de residus "ben alineats" era molt més reduït.

4.3. RESULTATS

PROTOCOL SCLQA

El problema d´evaluar la qualitat de les prediccions es divideix en dues parts: i) identificació del *fold* de la predicció i ii) comparació de la predicció amb homòlegs del mateix *fold* (amb la conseqüent extracció dels residus alineats). Per simplificar assumim que el primer pas es pot resoldre; de fet hi ha estudis fets al laboratori on s´ha aconseguit determinar el *fold* d´un conjunt de prediccions utilitzant mètodes de comparació estructural; també com s´ha vist al capítol 3, emprant sistemes d´intel.ligència artificial és possible assignar el *fold* a un conjunt de prediccions amb un rendiment suficient (aquest punt serà tractat amb més detall a l´apartat de Limitacions de SCLQA).

Figura 4.2. Esquema del protocol SCLQA. Es basa en alinear el model *de novo* amb un homòleg. S´assumeix que els residus alineats corresponen a aquelles parts del model de millor qualitat.



SCLQA es centra en el segon punt, i es basa en alinear el model *de novo* amb el/els homòlegs que tinguin el seu mateix *fold* (figura 4.2). Fet això assumim que els residus que han estat alineats són els corresponents a les parts estructurals del model amb millor qualitat. Cal remarcar que aquest mètode no utilitza un conjunt de prediccions, sinó que treballa sobre prediccions individuals.

CARACTERITZACIÓ DE LES PARTS ALINEADES

Les parts alineades amb els mètodes de comparació estructural o STR (Selected Target's Residues) s´han caracteritzat a dos nivells:

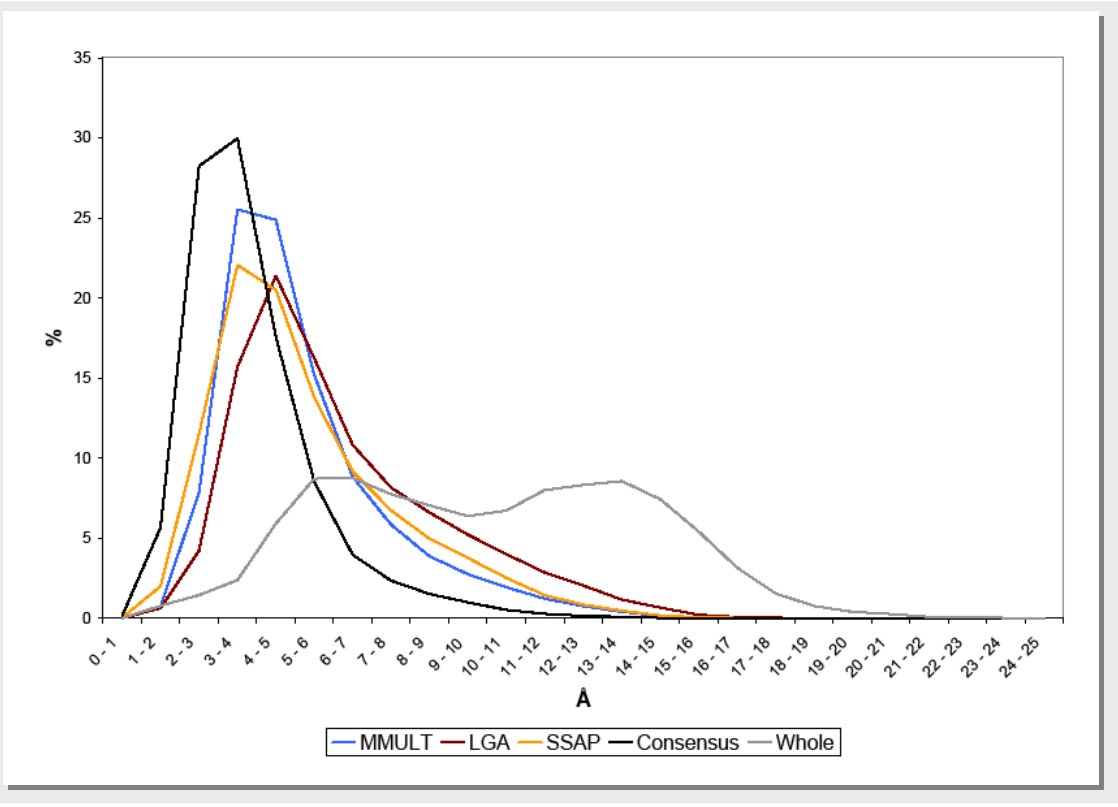
- Qualitat estructural: la qualitat estructural dels STR s´ha mesurat emprant dues variables, RMSD i GDT_TS. La primera per ser una mesura àmpliament utilitzada en la comparació estructural, i la segona per ser una mesura de qualitat molt usada en les avaluacions d´experiments CASP [1, 2]. Els RMSD i GDT_TS dels STR han estat comparats amb els valors anàlegs obtinguts quan s´utilitzen tots els residus de la predicció (ATR); d´aquesta manera es posa de manifest si la qualitat de les parts alineades és millor a la qualitat total de la predicció.
- Distribució al llarg de la seqüència: la distribució dels STR al llarg de la seqüència de la predicció s´ha caracteritzat utilitzant els tres paràmetres descrits a la part de *Metodologia*: distància màxima entre residus alineats (MD), mida del fragment alineat més gran (SLS) i distribució de mides dels fragments alineats (SAS).

Cal notar que per cada predicció/homòleg tindrem tres mètodes de comparació, el que donarà lloc a tres conjunts de STR ($STR_{MAMMOTH}$, STR_{LGA} i STR_{SSAP}). Pel que fa al conjunt ATR (All Target's Residues), sempre serà el mateix, independentment del mètode utilitzat (és la comparació de la predicció *de novo* amb l´estructura experimental).

RMSD

Els valors de RMSD es mostren a les figures 4.3 i 4.4. S´ha de tenir en compte que aquests valors corresponen a la superposició del model *de novo* i l´estructura experimental utilitzant el conjunt de residus STR i ATR. Això dóna lloc a l´RMSD de la part alineada i l´RMSD global respectivament. És important veure doncs que els valors d´RMSD no corresponen a la comparació entre la predicció i l´homòleg; aquesta comparació només serveix per obtenir el conjunt STR.

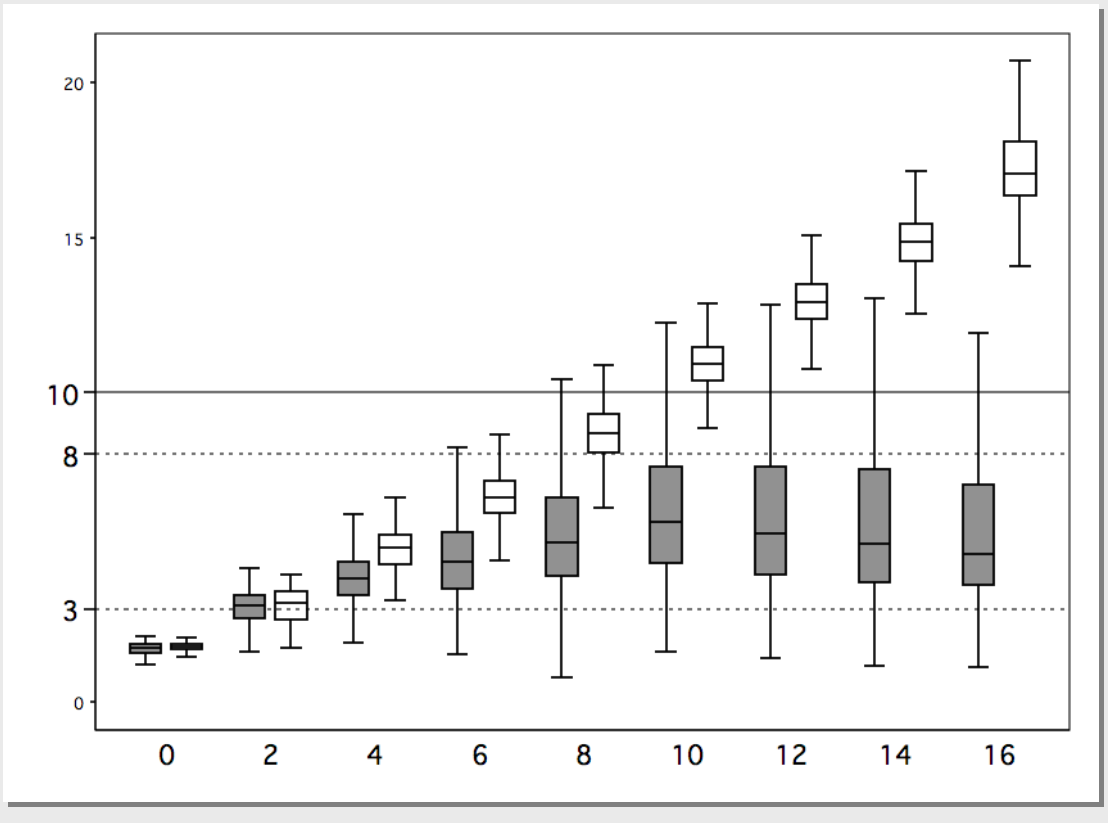
Figura 4.3. Distribucions d´RMSD per als conjunts STR (Selected Target Residues) corresponents a MAMMOTH, LGA, SSAP i Consensus. En gris es mostra la distribució d´RMSD corresponent a ATR (All Target Residues).



En la figura 4.3 es comparen les distribucions d´RMSD dels conjunts $STR_{MAMMOTH}$, STR_{LGA} i STR_{SSAP} amb la distribució de ATR. Es pot apreciar que independentment del mètode emprat les distribucions de STR presenten un desplaçament cap a valors de RMSD menors. Això indica que en general, els residus identificats per SCLQA presenten una major qualitat estructural. A la figura 4.5 s´il.lustra aquest punt amb un exemple,

l´acilfosfatasa bovina (pdb: 2acy): l´RMSD global de la predicció és 13.9Å, mentre que el de la part alineada baixa fins a 5.5Å. De fet la inspecció visual mostra una bona correspondència entre predicció i forma nativa. Tornant als resultats d´RMSD, és interessant notar que les distribucions de STR tenen un màxim al voltant de 5Å, i que la gran majoria es troba per sota dels 10Å.

Figura 4.4. Comparació dels RMSD de les parts seleccionades (eix y) i proteïna total (eix x). En gris es mostra la relació per als alineaments obtinguts amb MAMMOTH; el blanc les dades obtingudes a mode de control amb Prosa. La línia continua a 10Å actua de límit superior (la majoria de valors es troben per sota). Les línies discontinues delimiten l´àrea que conté un percentatge important de dades.



A la figura 4.4 es comparen la relació entre valors $RMSD_{STR}$ i $RMSD_{ATR}$, juntament amb els valors obtinguts amb Prosa (Veure *Metodologia*). Es poden distingir dues regions: per sota de 6-8Å a l´eix X existeix una relació lineal entre $RMSD_{ATR}$ i $RMSD_{STR}$; per sobre de 8Å s´assoleix un valor més aviat constant. Aquest comportament és el responsable del desplaçament de distribucions d´RMSD observat a la figura 4.4, i confirma la capacitat del mètode SCLQA per identificar conjunts de residus del model

amb millor qualitat que la resta. El valor constant es pot entendre com la contribució de dos factors:

- Distàncies màximes residu-residu imposades pels mètodes de comparació: els programes de comparació estructural imposen una distància màxima entre residus, per tal d´evitar aparellaments massa distants (per exemple a Mammoth i LGA són 4Å i 5Å, respectivament).
- Distàncies màximes residu-residu entre homòleg i nativa: les distàncies màximes entre els homòlegs i la nativa tenen un valor màxim concret, ja que comparteixen el mateix *fold*.

Si comparem els resultats de SCLQA amb els de Prosa, veiem que com passava en el cas anterior, per valors de ATR baixos existeix una relació força lineal entre RMSDs. No obstant això, a partir de 8Å aquesta tendència lineal es manté. Això demostra que SCLQA pot donar resultats que milloren/complementen els obtinguts per altres mètodes d´assignament de qualitats, sobretot quan les prediccions globalment són pobres.

Per tal de provar si es podia millorar el conjunt de STR, per cada model s´ha obtingut un conjunt STR consens, STR_{CONS}, que correspon a la intersecció entre $STR_{MAMMOTH}$, STR_{LGA} i STR_{SSAP}, és a dir, s´han agafat aquells residus de les prediccions que es trobaven presents a tots tres conjunts STR. A la figura 4.3 es mostra que realment els valors d´RMSD obtinguts emprant els STR_{CONS} són més baixos; no obstant això, les mides d´aquests conjunts de residus consens patien una reducció important (figura 4.6), a causa de les discrepàncies entre els tres mètodes d´alineament estructural; és per això que no han estat emprats en els anàlisis posteriors.

Figura 4.5. Acilfosfatasa bovina (pdb: 2acy). Al panell A trobem l´alineament entre el model *de novo* (blau) i l´homòleg 1vi7 (vermell). En el panell B trobem l´alineament entre el mateix model de novo (blau) i la seva forma nativa (taronja). Els alineaments han estat obtinguts amb MAMMOTH, i en ambdós casos els segments més gruixuts corresponen a les parts alineades.

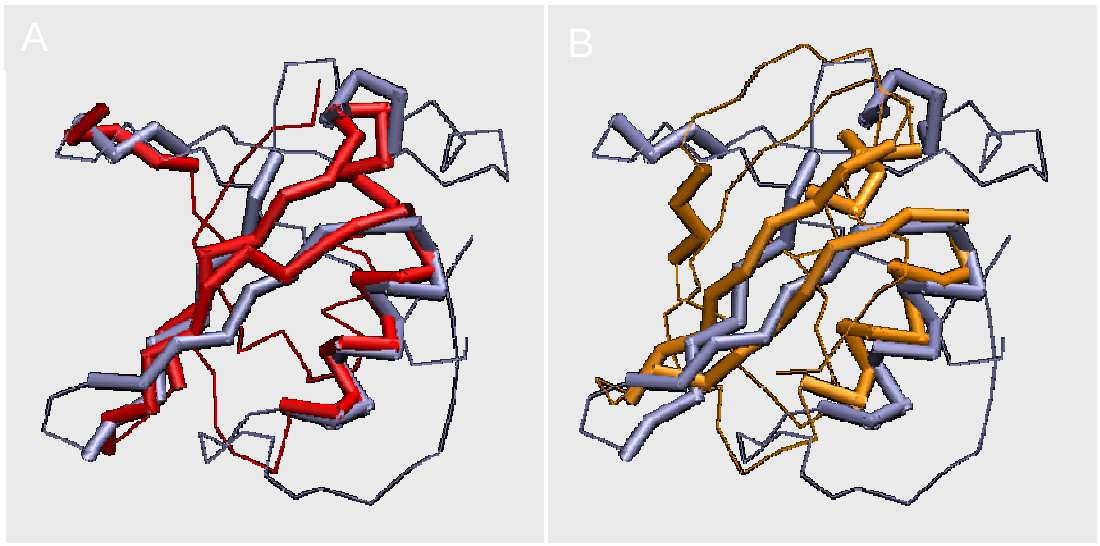
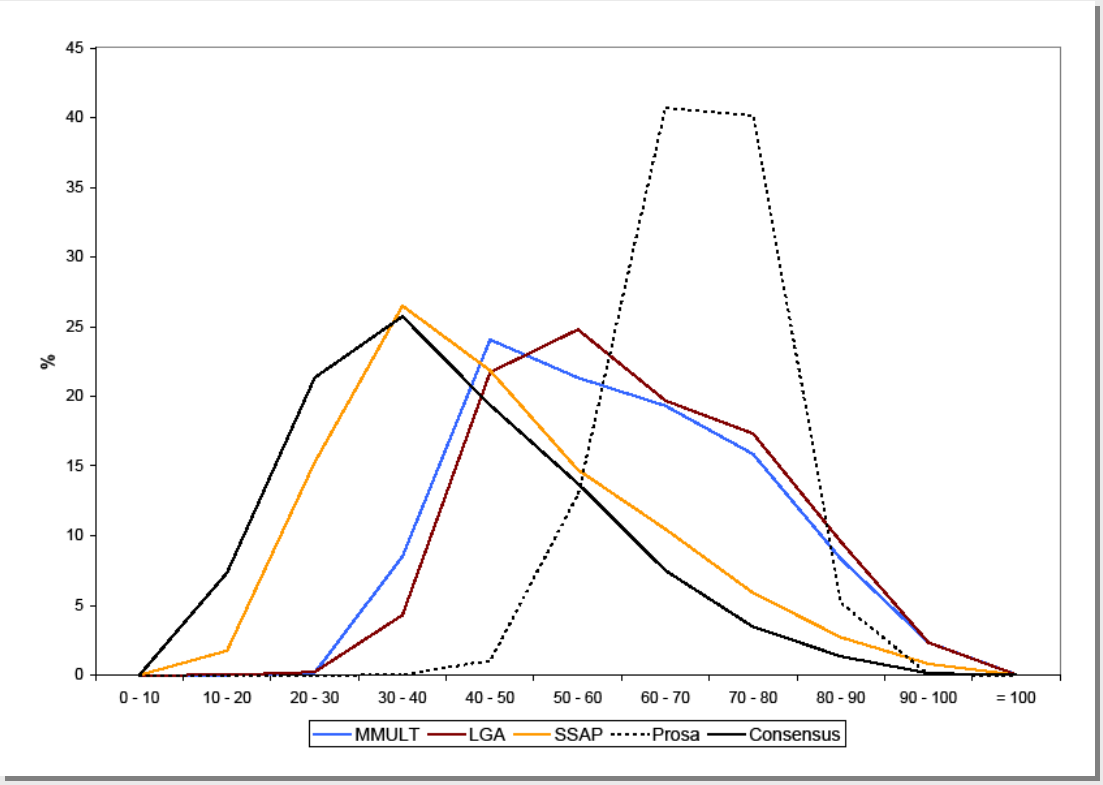


Figura 4.6. Percentatge de residus alineats (R) per a cadascun dels mètodes: MMULT, LGA i SSAP. En negre es presenten els valors per als STR consens, i en línia de punts els corresponents al control amb Prosa (veure metodologia).

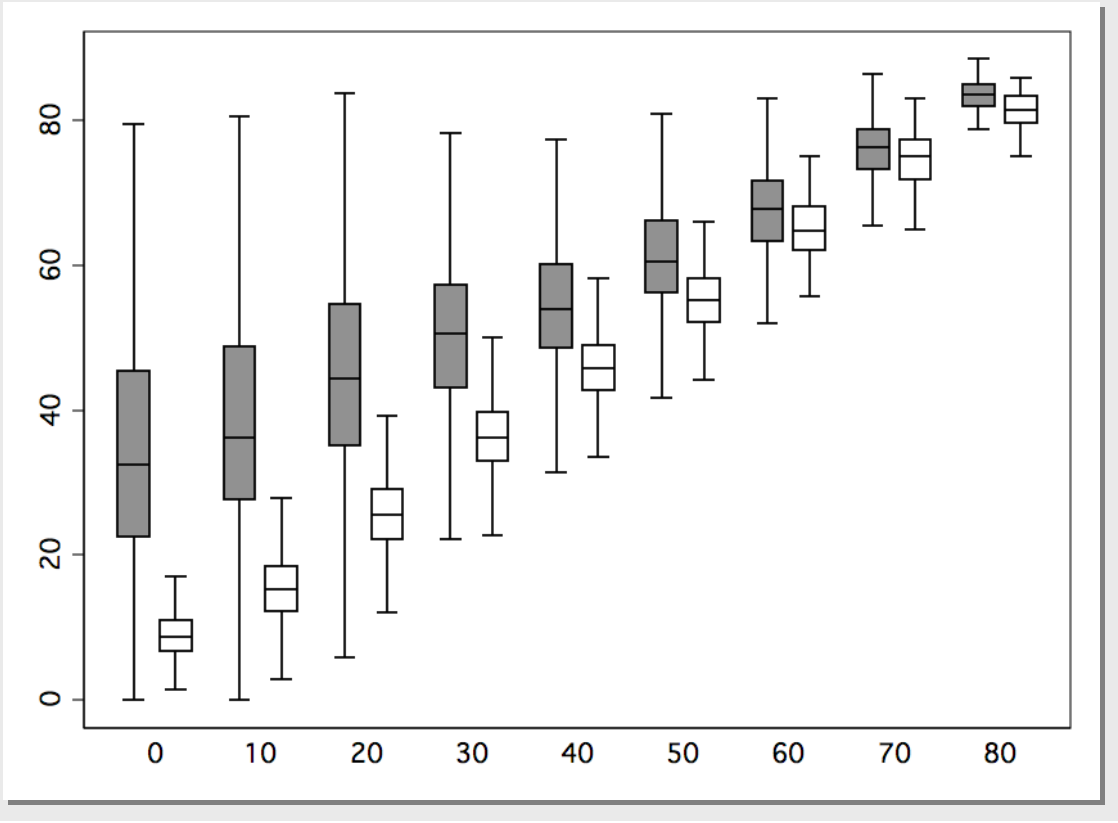


GDT_TS

GDT_TS permet detectar subestructures ben preservades. Valors alts propers a 100 indiquen la presència abundant de subestructures millor predites, mentre que valors baixos indiquen que les subestructures predites són poques.

Tal i com es pot veure a la figura 4.7, els valors de GDT_TS_{STR} són millors que els de GDT_TS_{ATR}. Aquesta tendència és particularment clara si ens fixem en la zona GDT_TS_{ATR} inferior a 40-50. Els resultats concorden amb els referents a l´RMSD, i confirmen que el protocol seguit permet seleccionar regions de les prediccions amb millor qualitat que la resta.

Figura 4.7. Comparació dels valors de GDT_TS per a les parts alineades (eix y) i tota la proteïna (eix x). En gris es presenten les dades utilitzant els alineaments estructurals obtinguts amb SSAP; en blanc i a mode de control, les dades obtingudes utilitzant Prosa.

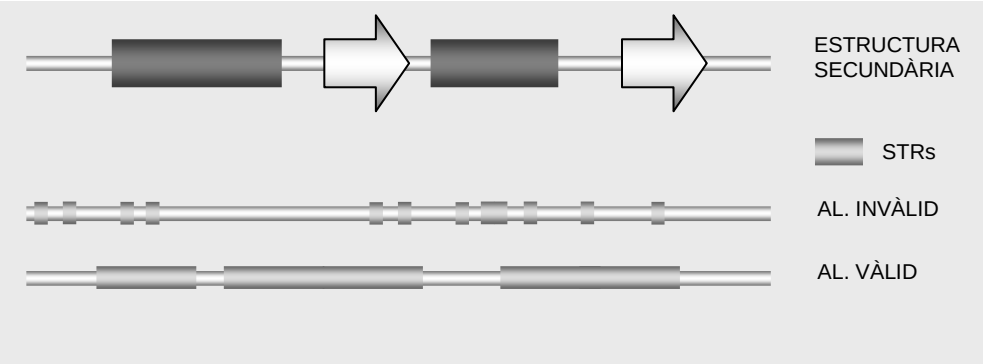


Si comparem els resultats de SCLQA amb els de Prosa veiem com tenen tendències similars per zones de GDT_TS_{ATR} alt; no obstant això, a mesura que la qualitat dels models baixa, SCLQA presenta millors resultats, i és capaç de trobar més zones de bona qualitat dins les prediccions. Això suggereix que per models pobres, l'ús de SCLQA pot ser una bona alternativa a altres mètodes d´assignament de qualitat.

DISTRIBUCIÓ DELS STR AL LLARG DE LA SEQÜÈNCIA

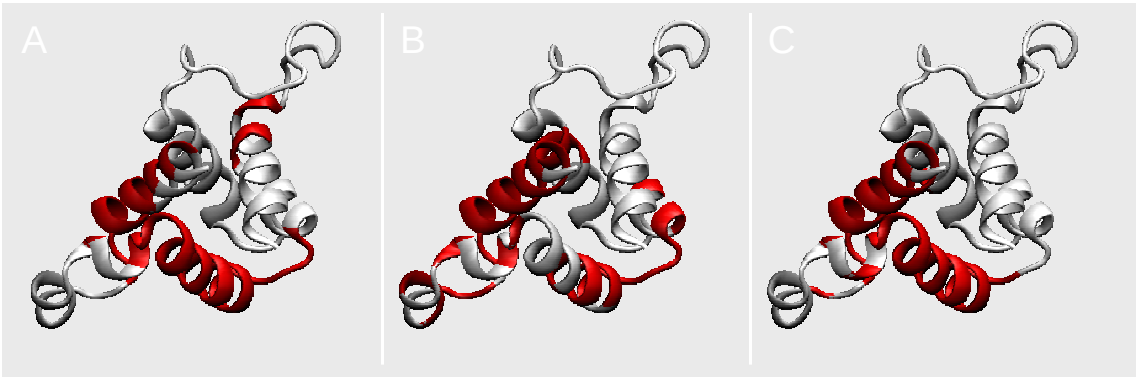
En els apartats anteriors hem assignat la qualitat estructural als STR; no obstant pot passar que tot i que la qualitat del conjunt de residus sigui bona, estructuralment no tinguin significat, per exemple per que els residus seleccionats es trobin molt dispersos al llarg de la seqüència... Aquests casos s´exemplifiquen al següent esquema:

Figura 4.8. Possibles distribucions dels STR al llarg de la seqüència.



En aquest apartat s´estudia la naturalesa d´aquests STR: com es distribueixen al llarg de la seqüència de la predicció. Tal i com es pot apreciar a la figura 4.9, la distribució dels conjunts de residus alineats és relativament heterogènia en tots tres mètodes de comparació estructural; ja que combinen fragments de diverses mides.

Figura 4.9. Distribució dels STR per al domini de Calponina de la beta-espectrina humana (pdb: 1aa2). En aquesta figura es mostren els residus alineats utilitzant SCQLA (vermell) per Mammoth (A), LGA (B) i SSAP (C).



Per tal de generalitzar l´estudi hem utilitzat tres mesures que donen una visió complementària de com es distribueixen els STR al llarg de la seqüència (*Metodologia*). Els nostres resultats mostren que els valors d´MD són majoritàriament superiors a 0.5; de fet més d´un 50% presenta valors superiors a 0.8 tal i com es pot veure a la figura 4.10A. Tots tres mètodes de comparació donen STR amb distribucions similars, tot i que la de LGA està lleugerament desplaçada cap a 1. La conclusió que es pot treure d´aquests resultats és que els STR es troben distribuïts al llarg de tota la seqüència de la predicció, per tant tenen una cobertura força àmplia.

Els resultats per SLS (figura 4.10B) i SAS (figura 4.10C) mostren que tot i que els STR estan formats per fragments de mida reduïda, ja que aproximadament un 50% d´aquests són menors al 10% de residus alineats, hi ha un percentatge important de fragments de mida major; per exemple més del 95% dels valors de SLS estan per sobre del 30%.

Resumint, el panorama que ens presenta SCLQA és que en general aquests residus amb millor qualitat es troben distribuïts al llarg de tota la seqüència. Cal remarcar però que els fragments d´aquests STR són força heterogenis, ja que presenten mides molt dispars; tot i que tenen mida suficient com per correspondre a elements estructurals bàsics com hèlix o fulles beta.

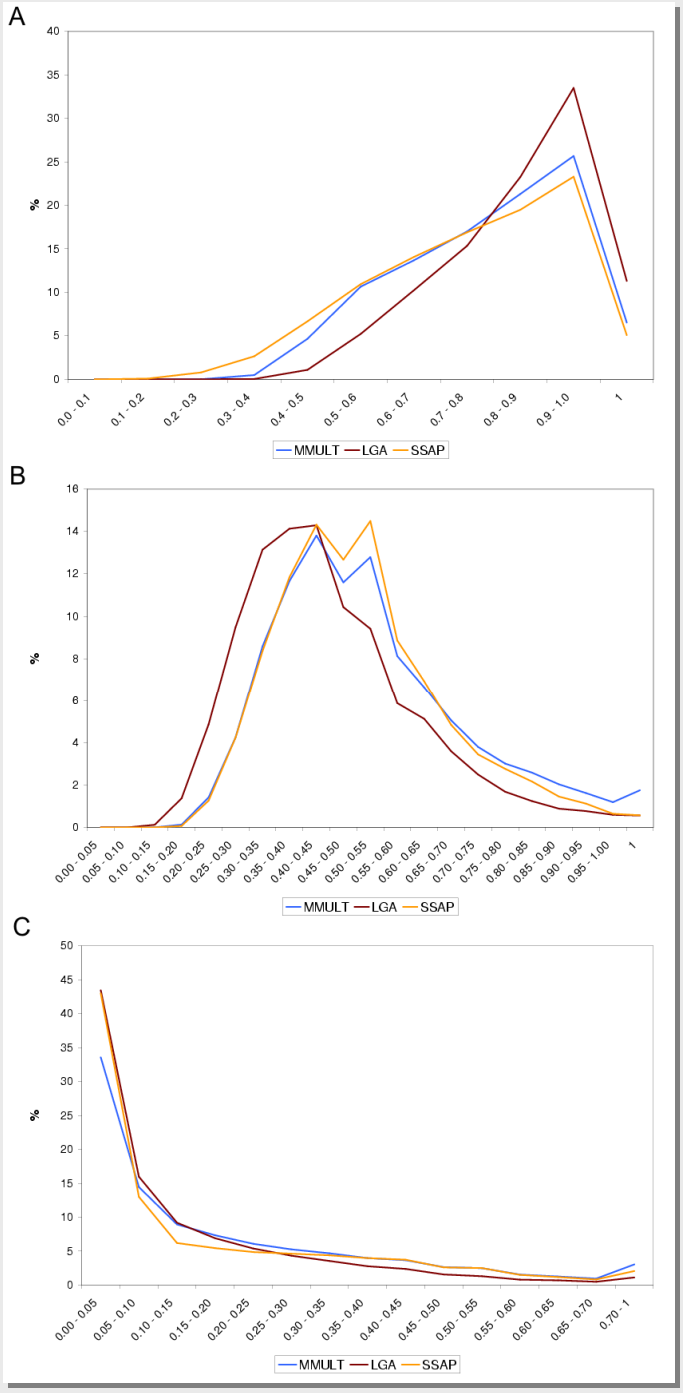


Figura 4.10. Distribució dels STR al llarg de la seqüència. Es mostren les distribucions de MD, SLS i SAS als panells A, B i C respectivament.

MILLORA DE LA QUALITAT DELS STR

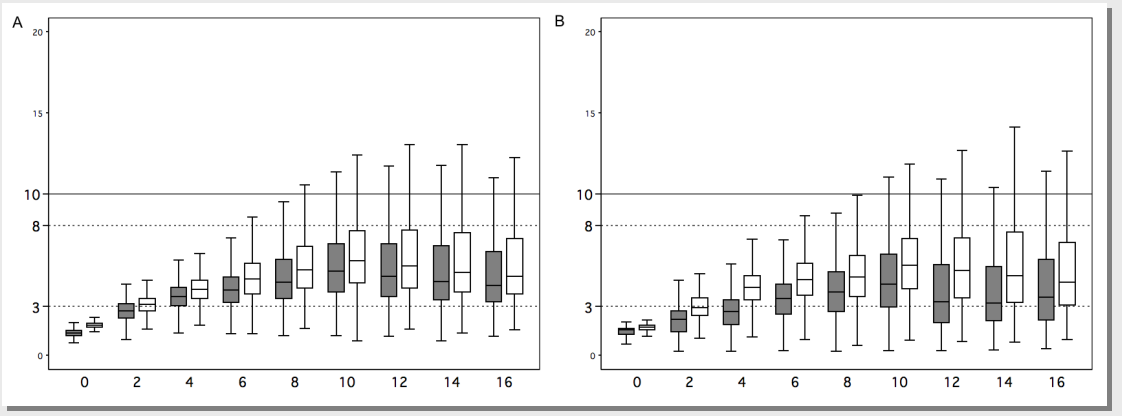
Hem vist com agafant els residus consens (STR_{CONS}) som capaços de millorar el conjunt de residus seleccionats per SCLQA. En base a aquest fet ens hem plantejat si és possible utilitzar l´accessibilitat i l´estructura secundària de forma similar, és a dir,

comprovar si els residus més/menys accessibles, o amb/sense estructura secundària presenten una millor qualitat dins el conjunt de STR. Més concretament, en el cas de l´accessibilitat hem separat els STR entre residus accessibles i no accessibles (accessibilitat relativa superior i inferior 20% respectivament). En el cas de l´estructura secundària hem classificat els STR segons si presenten una estructura periòdica (hèlix o fulla beta) o no periòdica. Cal remarcar que els càlculs d´accessibilitat i les prediccions d´estructura secundària han estat fetes no sobre les prediccions, sinó sobre els homòlegs, i han estat assignats a les prediccions utilitzant la correspondència de residus predicció-homòleg dels alineaments estructurals.

En el cas de l´accessibilitat, els residus de les prediccions aparellats amb residus de l´homòleg enterrats presenten valors d´RMSD lleugerament més baixos que aquells aparellats amb residus accessibles (figura 4.11A). En el cas de l´estructura secundària passa quelcom similar, i els residus de les prediccions aparellats amb residus de l´homòleg amb estructura definida (hèlix o fulla beta) presenten RMSDs una mica menors (figura 4.11B). Té sentit si tenim en compte que tant la part central de les proteïnes, com els motius estructurals es troben molt més conservats, per tant variaran menys.

Aquests resultats suggereixen que aquests dos paràmetres poden ser útils, sobretot si es combinen per exemple amb l'ús d´alineaments consens; de totes maneres cal tenir en compte que aplicar aquests criteris de selecció comporten una reducció de la mida dels STR importants, fet que pot ser contraproduent, ja que aleshores l´espai conformacional dels residus restants pot arribar a ser molt extens, i el seu refinament esdevenir directament un problema de predicció de *folding*.

Figura 4.11. Obtenció de conjunts de residu de millor qualitat. Els STRs originals s´han partit segons dos criteris: accessibilitat i estructura secundària. Al panell A es mostra la distribució de RMSD per als residus enterrats (gris) i accessibles (blanc). Al panell B es mostra la distribució de RMSD per als residus amb estructura secundària ordenada (gris) i desordenada (blanc). El significat de les línies discontinues és el mateix que a la figura 4.4; de la mateixa manera l´eix X representa l´RMSD global de les prediccions. Els alineaments utilitzats són de MAMMOTH.



LÍMITS DE SCLQA

En aquest apartat es comenten les limitacions del protocol desenvolupat. Les dues primeres deriven directament de les assumpcions del protocol (*Metodologia*).

- Les prediccions han de disposar d´homòlegs en les bases de dades estructurals. Aquesta limitació molt probablement trobi solució en un futur proper, gràcies a la gran cobertura estructural que estan assolint els projectes de genòmica estructural [29]. Estudis recents suggereixen que s´està aconseguint una cobertura completa de l´espai estructural de dominis [30].

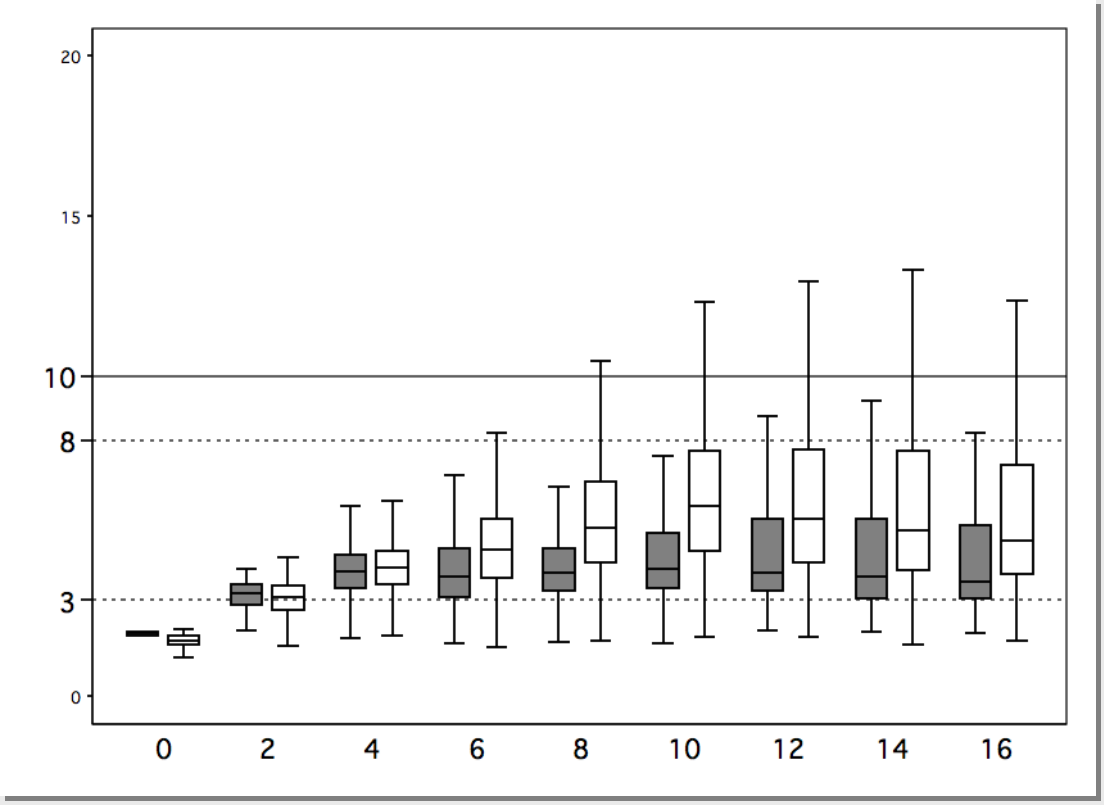
- Les prediccions han de presentar un *fold* correcte. Aquest punt deriva directament del funcionament dels mètodes de predicció estructural, els quals són incapaços de generar prediccions amb *fold* correcte de forma consistent. Aquesta limitació implica que abans d´aplicar SCLQA s´ha de verificar que el *fold* de la predicció sigui correcte. Tot i que no és un problema trivial hi ha diverses opcions a seguir. Una possibilitat seria puntuar la qualitat global de la predicció emprant programes com Prosa o Verify3D, i seguidament identificar

quin és el *fold* a que pertany aquesta predicció. Per aquest darrer pas, els resultats mostrats al capítol 3, treballs fets al grup [16] i altres treballs realitzats per David Baker [15], mostren que utilitzant mètodes de comparació estructural es pot arribar a identificar el *fold* d´una predicció.

- Distància entre la proteïna a predir i els seus homòlegs. Si l´homòleg és distant a la proteïna a predir, els mètodes de comparació estructural poden imposar erròniament a la predicció motius estructurals presents únicament en l´homòleg. Per tal d´explorar aquesta possibilitat hem estudiat la contribució de la similitud de seqüència entre predicció i homòleg. Més concretament, hem dividit el conjunt de prediccions en dos conjunts, segons si la seva identitat de seqüència amb l´homòleg és inferior o superior al 30%. A la figura 4.12 es mostra la relació entre RMSD de les parts alineades (STR) i RMSD global per als dos conjunts. Tal i com es podia esperar, les parelles amb identitats per sobre del 30% presenten RMSDs menors que aquelles amb identitats inferiors al 30%; no obstant això, fins i tot en aquestes darreres veiem com hi ha una millora de qualitat dels STR respecte la qualitat global de la predicció, confirmant així que SCLQA pot ser útil també quan només es disposen d´homòlegs distants.

- SCLQA no proporciona una puntuació específica per cada residu, únicament els separa en bons i dolents, existint la possibilitat de refinar aquesta classificació utilitzant STR consens, informació d´accessibilitat o d´estructura secundària. No obstant això un sistema de puntuació de la qualitat de cada residu continua essent necessari. Una possible solució seria emprar programes com Prosa [8], Verify3D [10], o el conjunt de programes desenvolupats per Arne Elofsson [14, 31, 32].

Figura 4.12. Efecte de la distància proteïna-homòleg. Es mostren les dades de la figura 4.4 partides segons la identitat de seqüència entre la proteïna i l´homòleg. Es mostren en blanc i gris les dades de les comparacions amb identitat de seqüència és inferior o superior al 30% respectivament.



4.4. DISCUSSIÓ

Els resultats presentats en aquest capítol mostren que és possible utilitzar mètodes de comparació estructural per classificar els residus de les prediccions *de novo* en dues categories des d´un punt de vista de qualitat, fins i tot quan només es disposa d´homòlegs llunyans a la predicció. El conjunt de residus de bona qualitat pot ser fins i tot millorat moderadament si introduïm variables com l´accessibilitat dels residus, o la propensió a formar elements d´estructura secundària.

Els principals avantatges del mètode presentat són per una banda la seva simplicitat conceptual i d´aplicació. Els principals inconvenients són que el mètode es basa dues assumpcions:

- Els models han de presentar un *fold* correcte.
- Hem de disposar d´homòlegs a les prediccions.

No obstant això, els mètodes d´assignació de qualitat milloren cada dia, i els programes de genòmica estructural cada vegada cobreixen un major espai estructural; així doncs és d´esperar que aquests dos punts en un futur no gaire llunyà deixaran de ser un problema.

Pel que fa a l´aplicabilitat, en el següent capítol es presenta un protocol de refinament de prediccions *de novo* basat dels resultats presentats.

4.5. REFERÈNCIES

1. Jauch, R., et al., *Assessment of CASP7 structure predictions for template free targets.* Proteins, 2007. **69 Suppl 8**: p. 57-67.
2. Vincent, J.J., et al., *Assessment of CASP6 predictions for new and nearly new fold targets.* Proteins, 2005. **61 Suppl 7**: p. 67-83.
3. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
4. Laskowski, R.A., M.W. MacArthur, and J.M. Thornton, *Validation of protein models derived from experiment.* Curr Opin Struct Biol, 1998. **8**(5): p. 631-9.
5. Kleywegt, G.J., *Validation of protein crystal structures.* Acta Crystallogr D Biol Crystallogr, 2000. **56**(Pt 3): p. 249-65.
6. Laskowski, R.A., et al., *PROCHECK: a program to check the stereochemical quality of protein structures.* J Appl Crystallogr, 1993. **26**: p. 283-291.
7. Hooft, R.W., C. Sander, and G. Vriend, *Objectively judging the quality of a protein structure from a Ramachandran plot.* Comput Appl Biosci, 1997. **13**(4): p. 425-30.
8. Sippl, M.J., *Recognition of errors in three-dimensional structures of proteins.* Proteins, 1993. **17**(4): p. 355-62.
9. Melo, F., et al., *ANOLEA: a www server to assess protein structures.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 187-90.
10. Luthy, R., J.U. Bowie, and D. Eisenberg, *Assessment of protein models with three-dimensional profiles.* Nature, 1992. **356**(6364): p. 83-5.
11. Vriend, G., *WHAT IF: a molecular modeling and drug design program.* J Mol Graph, 1990. **8**(1): p. 52-6, 29.
12. Lundstrom, J., et al., *Pcons: a neural-network-based consensus predictor that improves fold recognition.* Protein Sci, 2001. **10**(11): p. 2354-62.
13. Wallner, B. and A. Elofsson, *Pcons5: combining consensus, structural evaluation and fold recognition scores.* Bioinformatics, 2005. **21**(23): p. 4248-54.

14. Wallner, B. and A. Elofsson, *Prediction of global and local model quality in CASP7 using Pcons and ProQ.* Proteins, 2007. **69 Suppl 8**: p. 184-93.

15. Bonneau, R., et al., *De novo prediction of three-dimensional structures for major protein families.* J Mol Biol, 2002. **322**(1): p. 65-78.

16. de la Cruz, X., I. Sillitoe, and C. Orengo, *Use of structure comparison methods for the refinement of protein structure predictions. I. Identifying the structural family of a protein from low-resolution models.* Proteins, 2002. **46**(1): p. 72-84.

17. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

18. Aloy, P., et al., *Predictions without templates: new folds, secondary structure, and contacts in CASP5.* Proteins, 2003. **53 Suppl 6**: p. 436-56.

19. Lesk, A.M., L. Lo Conte, and T.J. Hubbard, *Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts.* Proteins, 2001. **Suppl 5**: p. 98-118.

20. Orengo, C.A., et al., *Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction.* Proteins, 1999. **Suppl 3**: p. 149-70.

21. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.

22. Pearl, F., et al., *The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.* Nucleic Acids Res, 2005. **33**(Database issue): p. D247-51.

23. Orengo, C.A. and W.R. Taylor, *A rapid method of protein structure alignment.* J Theor Biol, 1990. **147**(4): p. 517-51.

24. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.

25. Kabsch, W.A., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallogr A, 1976. **A32**: p. 922-923.

26. Hubbard, S.J. and J. Thornton, *NACCESS.* Department of biochemistry and Molecular Biology, University College Londo, 1993.

27. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

28. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

29. Todd, A.E., et al., *Progress of structural genomics initiatives: an analysis of solved target structures.* J Mol Biol, 2005. **348**(5): p. 1235-60.

30. Kihara, D. and J. Skolnick, *The PDB is a covering set of small protein structures.* J Mol Biol, 2003. **334**(4): p. 793-802.

31. Wallner, B. and A. Elofsson, *Can correct protein models be identified?* Protein Sci, 2003. **12**(5): p. 1073-86.

32. Wallner, B. and A. Elofsson, *Identification of correct regions in protein models using structural, alignment, and consensus information.* Protein Sci, 2006. **15**(4): p. 900-13.

CAPÍTOL 5.

REFINAT DE PREDICCIONS

DE NOVO

5.1. INTRODUCCIÓ

En els darrers anys el camp de la predicció estructural ha fet un gran progrés, i avui dia és possible generar prediccions emprant diversos mètodes, que enfoquen el problema de forma diferent, com per exemple mètodes de predicció *de novo* –intenten predir l´estructura a partir de la seqüència basant-se en principis físics del plegament de proteïnes- o mètodes basats en el modelat per homologia [1-4] –intenten predir l´estructura d´una seqüència a partir de la seva similitud amb una altra d´estructura coneguda. No obstant això, la qualitat de les prediccions mirada de forma global no és prou bona com per a ser utilitzades en aplicacions que requereixin un detall estructural elevat (ex. disseny de fàrmacs).

El principal motiu d´aquesta "manca de precisió" la trobem en les aproximacions utilitzades pels programes de predicció [5]. En les tècniques de predicció *de novo* les cadenes laterals es solen representar de forma simplificada, sense hidròdens, i només amb certes geometries possibles (restricció d´angles torsionals per exemple), en conseqüència les interaccions entre els diferents aminoàcids passen a ser també aproximades, etc. En el cas del modelat per homologia un dels punts crítics és el problema de predicció dels *loops*, normalment situats a zones menys conservades i per tant pitjor alineades [6].

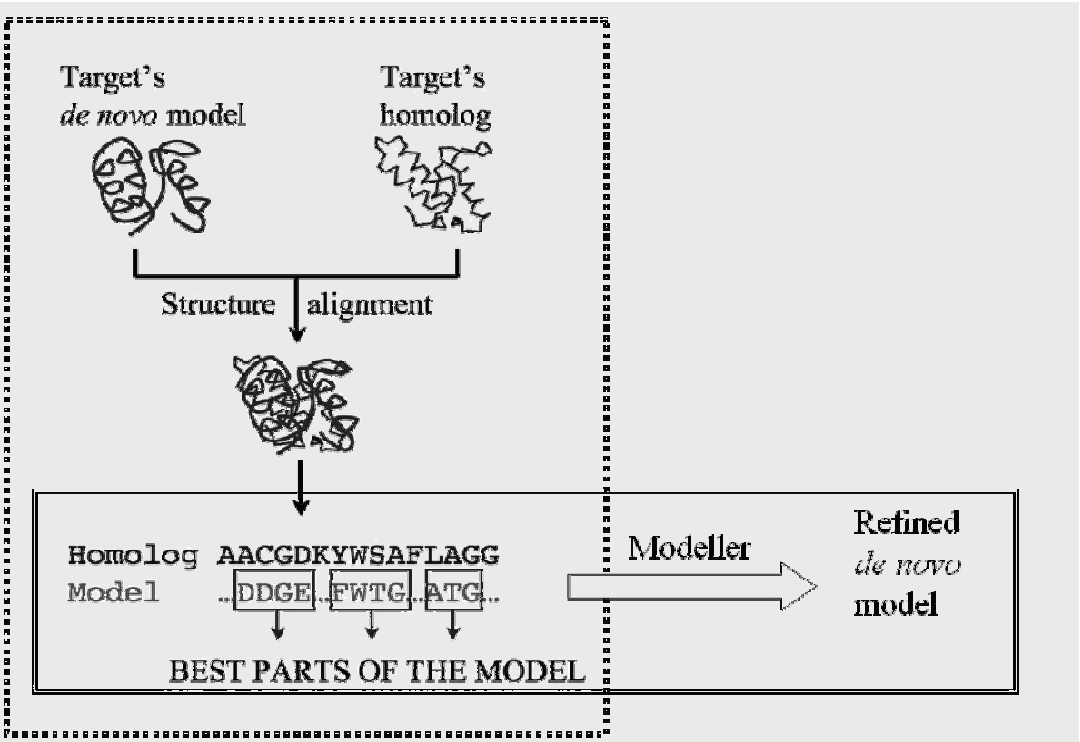
Per tal de solucionar aquests problemes cada cop més s´està plantejant com alternativa el desenvolupament de mètodes de refinament [7-9]. D´aquesta manera el problema de predicció estructural queda dividit en dos parts: obtenció d´una estructura de baixa resolució i refinament de la mateixa. Aquesta estratègia resulta bastant raonable, ja que permet utilitzar en la segona etapa mètodes computacionalment més costosos, ja que l´espai conformacional a explorar és menor. No obstant això, el problema continua essent complicat, ja que per una banda es requereix una tècnica de mostreig efectiva, i per l´altra és completament necessari

disposar de funcions d´energia acurades que permetin diferenciar entre aquelles prediccions correctes i les que no ho són.

Cal dir però, que tot i aquestes dificultats s´han desenvolupat diverses estratègies encaminades a refinar les prediccions. Així per exemple, Lee i col.laboradors [10] emprant dinàmiques moleculars han estat capaços de distingir entre prediccions estructuralment similars a la proteïna nativa i prediccions amb estructures incorrectes. En altres estudis duts a terme per David Baker [11, 12] s´han desenvolupat funcions d´energia optimitzades per a tal finalitat. Pel que fa als models per homologia és comú utilitzar mètodes de refinament de *loops* [13-21] i de cadenes laterals [22] o fins i tot dinàmiques moleculars a escales de temps llargues [23], tot i que en aquest cas el punt principal de refinament continua essent millorar l´alineament (veure *Introducció*). En qualsevol cas, resulta cada cop més evident que una estratègia de refinat passa abans per la identificació de les zones de diferent qualitat en les prediccions, per procedir a la seva posterior millora [24].

En el capítol anterior s´ha mostrat com utilitzant un senzill protocol (SCQLA) és possible definir quines són les parts de millor qualitat en un conjunt de prediccions *de novo* generades amb Rosetta [25]. En el capítol que ens ocupa ampliem aquest estudi i mostrem com utilitzar aquesta informació per tal de refinar la qualitat inicial de les prediccions *de novo* utilitzades. Concretament volem comprovar si sóm capaços de refinar aquestes prediccions utilitzant la informació sobre zones de bona qualitat de l´estudi anterior i tècniques de modelat per homologia. Al següent esquema es pot veure resumit el procés:

Figura 5.1. Esquema del protocol de refinat. Es basa en els alineaments estructurals obtinguts en el capítol anterior. Aquests alineaments són utilitzats per construir models per homologia de les prediccions *de novo* originals.



5.2. METODOLOGIA

CONJUNT DE PREDICCIONS

De les 68 proteïnes estudiades en el capítol anterior n´hem seleccionat 15, que cobreixen les tres classes CATH [26]: alfa, beta i alfa/beta. El total de prediccions *de novo* involucrades en l´estudi és de 2693. A la següent taula es pot apreciar la contribució de cada proteïna al total de prediccions *de novo*:

Taula 5.1. Proteïnes utilitzades en el treball, clasificades segons la seva classe CATH (Alfa, Beta, Alfa/Beta), juntament amb el número de prediccions *de novo* amb que contribueix cadascuna.

Proteïna	# prediccions Alfa	Proteïna	# prediccions Beta	Proteïna	# prediccions Alfa/Beta
1aa2	84	1csp	207	1ctf	388
1ag2	31	1fbr	27	1fwp	105
1ail	182	1bdo	25	1aa3	99
1c5a	193	1ark	252	1svq	230
1hsn	468	1tit	234	1acf	168

ALINEAMENTS ESTRUCTURALS

En el capítol anterior hem vist com utilitzant programes de comparació estructural érem capaços de definir les regions de més bona qualitat en conjunt de prediccions *de novo*. Hem vist també que aquestes regions corresponen a les parts de la predicció que han estat alineades amb el corresponent homòleg.

Tal i com s´ha comentat en aquest capítol pretenem construir un conjunt de models per homologia emprant els alineaments estructurals obtinguts en el capítol anterior. Per fer-ho utilitzarem un total de 8033 d´aquests alineaments obtinguts de les corresponents comparacions dutes a terme amb Mammoth [27].

MODELAT PER HOMOLOGIA

S´ha utilitzat el programa Modeller [28], executat amb paràmetres estàndards, però utilitzant com a entrada els 8033 alineaments estructurals entre els models *de novo* de la proteïna d´interès i el seu homòleg a CATH (mirar capítol anterior).

El resultat ha estat un total de 8010 models (la diferència es deu a que el programa de modelatge ha fallat en alguns casos on hi havia discrepàncies entre la seqüència i el *target* utilitzat).

CÀLCUL D´RMSD I GDT_TS

Per tal d´avaluar la millora dels models obtinguts respecte les prediccions *de novo* inicials hem utilitzat dos paràmetres: RMSD [29] i GDT_TS [30]. Per un major detall sobre aquestes mesures mirar l´apartat de *Metodologia general*.

5.3. RESULTATS

Tal i com s´ha descrit a la metodologia, l´objectiu del present capítol és contrastar un senzill protocol de refinament estructural basat en la construcció d´un conjunt de models per homologia utilitzant alineaments estructurals. Per tal d´avaluar l´èxit d´aquesta estratègia comparem els models resultants amb les prediccions *de novo* originals. A les figures 5.2A i 5.2B es veu una primera comparació entre models per homologia i prediccions *de novo* a nivell d´RMSD: es presenten les distribucions d´RMSD globals i de les zones alineades respectivament, per a les prediccions *de novo* (en blanc) i per als models per homologia generats després del refinament (en gris). Com es pot apreciar, en tots dos casos la distribució d´RMSD dels models refinats es troba desplaçada cap a valors més baixos, indicant per tant un guany en la qualitat global i en la qualitat de les parts alineades.

Per tal d´il.lustrar aquests resultats de forma més específica mostrem la figura 5.3, on podem veure el cas del domini calponina de la beta-espectrina humana (PDB: 1aa2): en blanc es representa la proteïna nativa; en gris el model per homologia refinat, i en negre la predicció *de novo* inicial. La inspecció visual mostra que l´empaquetament i plegament del model per homologia és més proper a l´estructura nativa que el corresponent a la predicció *de novo*, corroborant per tant els resultats mostrats als histogrames de les figures 5.2A i 5.2B.

Figura 5.2. Distribucions d´RMSD considerant la proteïna sencera (A) i només les parts alineades (B). En gris es presenten els valors corresponents als models refinats; en blanc els corresponents a les prediccions *de novo* inicials.

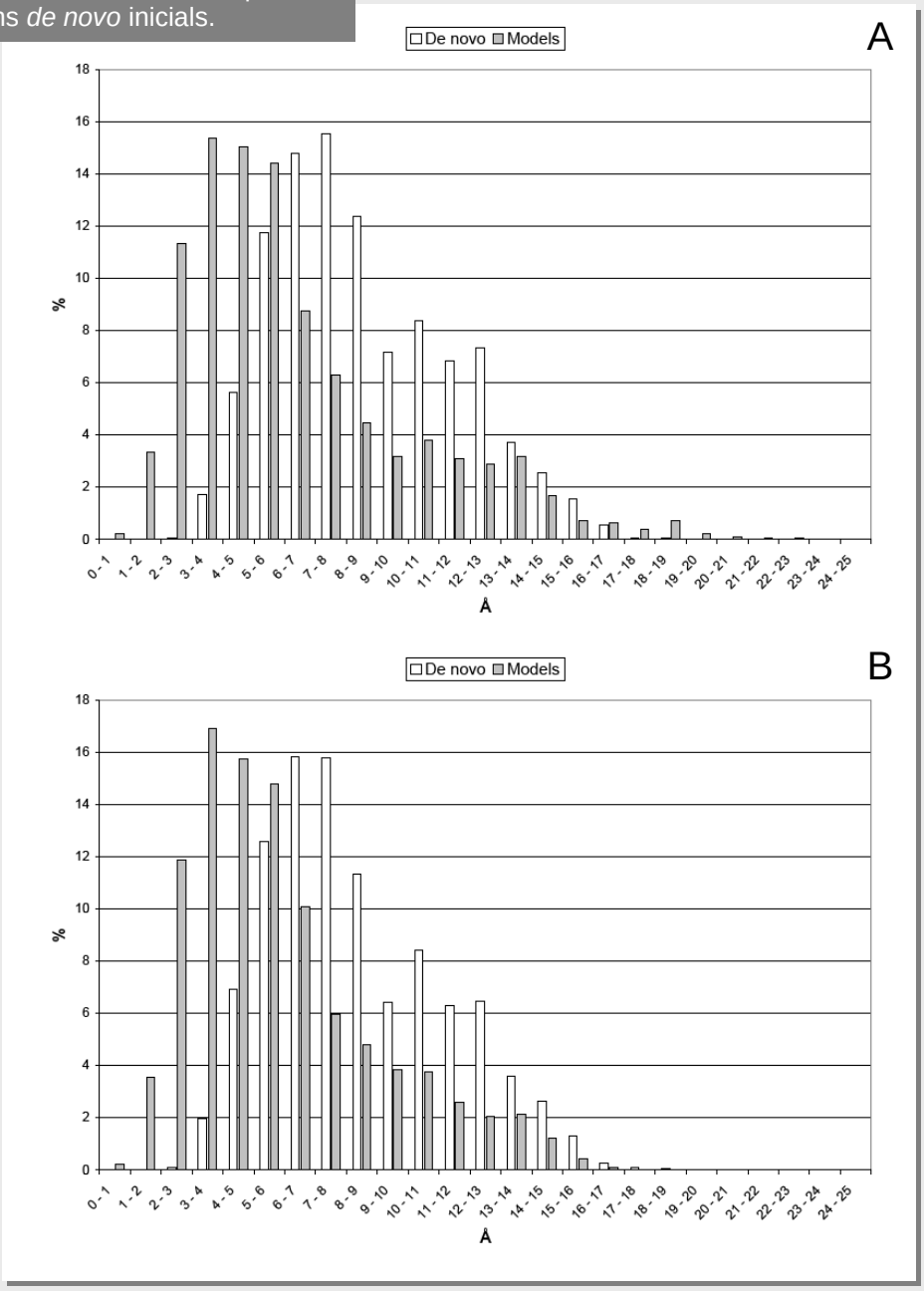
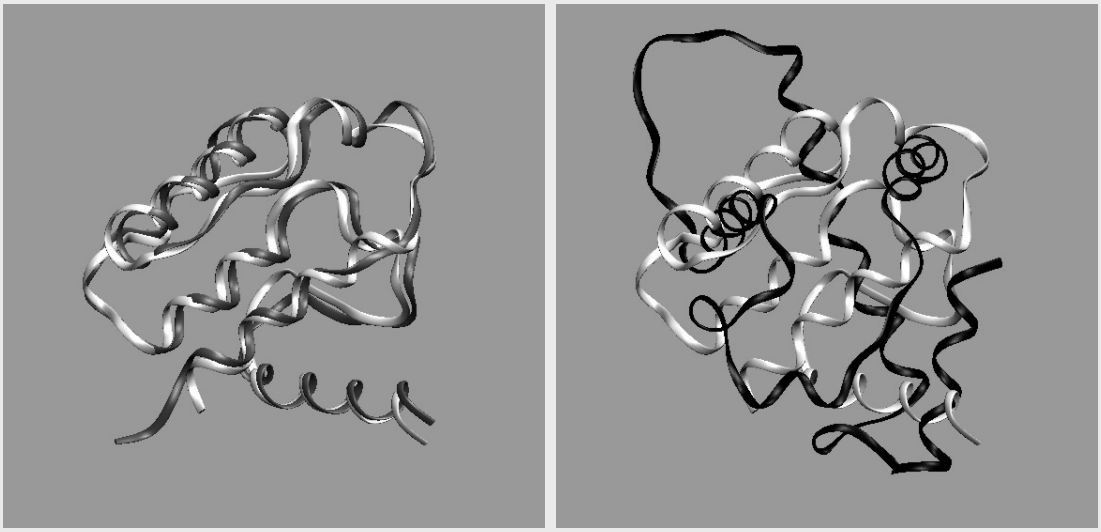
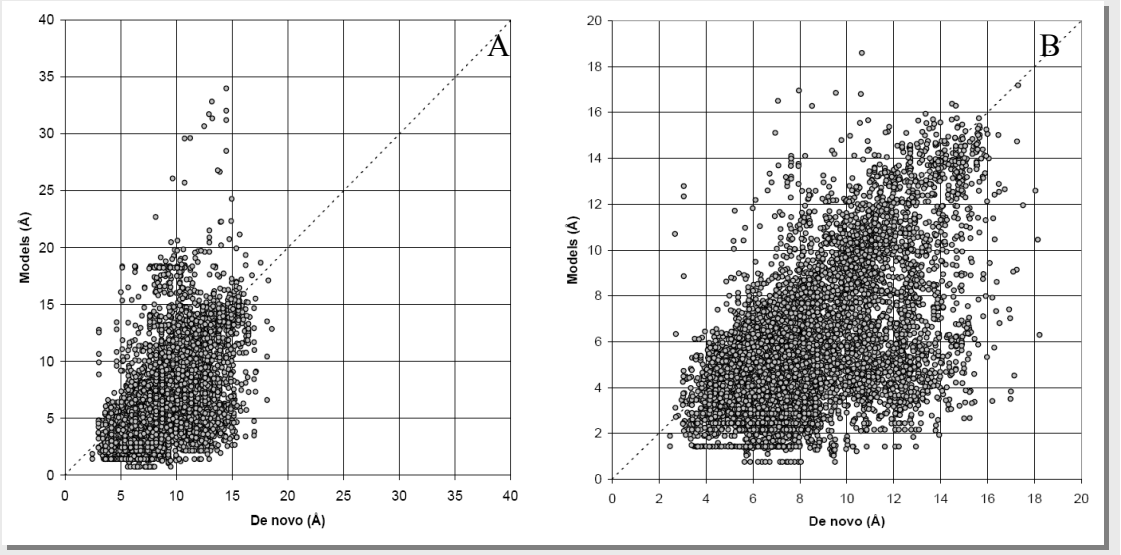


Figura 5.3. Domini calponina de la beta-espectrina humana (PDB: 1aa2): en blanc es representa la proteïna nativa; en gris (esquerra) el model per homologia refinat, i en negre (dreta) la predicció *de novo* inicial



De forma més detallada a les figures 5.4A i 5.4B es mostra la relació entre RMSD global i RMSD de la part alineada per cada parella predicció *de novo*/model refinat; això permet veure la variació real de cada predicció. Tot i que hi ha certs casos on els valors d´RMSD dels models refinats (eix y) es troben per sobre la diagonal, i per tant indiquen que la seva qualitat és inferior a la de les prediccions *de novo* equivalents, la gran majoria de punts mostren que de forma general la qualitat dels models refinats és superior. Requereixen una menció especial alguns models refinats amb RMSD global propers a 30Å (figura 5.4A), ja que s´escapen de forma considerable de la tendència general; la raó d´aquest comportament es pot explicar mirant la figura 5.6: existeix una relació inversa entre RMSD i fracció de residus alineats, per tant aquells models amb RMSD elevats deuen la seva baixa qualitat a que han estat construïts amb alineaments dolents.

Figura 5.4. A la figura A es mostra la relació entre RMSD global de les prediccions *de novo* inicials (eix x) i els models refinats (eix y). A la figura B es mostra la mateixa relació però considerant l´RMSD de les parts alineades.



A part d´utilitzar RMSD com a variable de comparació d´estructures, és molt comú en els àmbits de la predicció estructural utilitzar el GDT_TS [30]; aquesta mesura aporta una visió sobre l´existència de subestructures de qualitat (mirar *Metodologia general)*. A les figures 5.5A i 5.5B podem veure la relació de valors GDT_TS global i de la zona alineada respectivament, per als parells predicció *de novo*/model refinat. Cal tenir present que a diferència del RMSD, valors elevats de GDT_TS indiquen alta similitud estructural. Així doncs en el cas del GDT, tal i com succeïa amb l´RMSD, els models refinats presenten valors majors a les prediccions *de novo* (la gran majoria de punts es troben per sobre de la diagonal), indicant per tant una millor qualitat. Cal comentar també que en el cas del GDT l´efecte de la fracció de residus alineats no sembla jugar un paper tant important com el l´RMSD (figura 5.6); la raó pot ser que a diferència de l´RMSD, que es basa en la distància estructural entre parells, el GDT_TS treballa buscant subestructures dins aquests conjunts de punts, per tant el fet que hi hagi residus o parts del model per homologia no alineats pot tenir menys repercussió negativa sobre la qualitat global.

Figura 5.5. A la figura A es mostra la relació entre GDT_TS global de les prediccions *de novo* inicials (eix x) i els models refinats (eix y). A la figura B es mostra la mateixa relació però considerant el GDT_TS de les parts alineades.

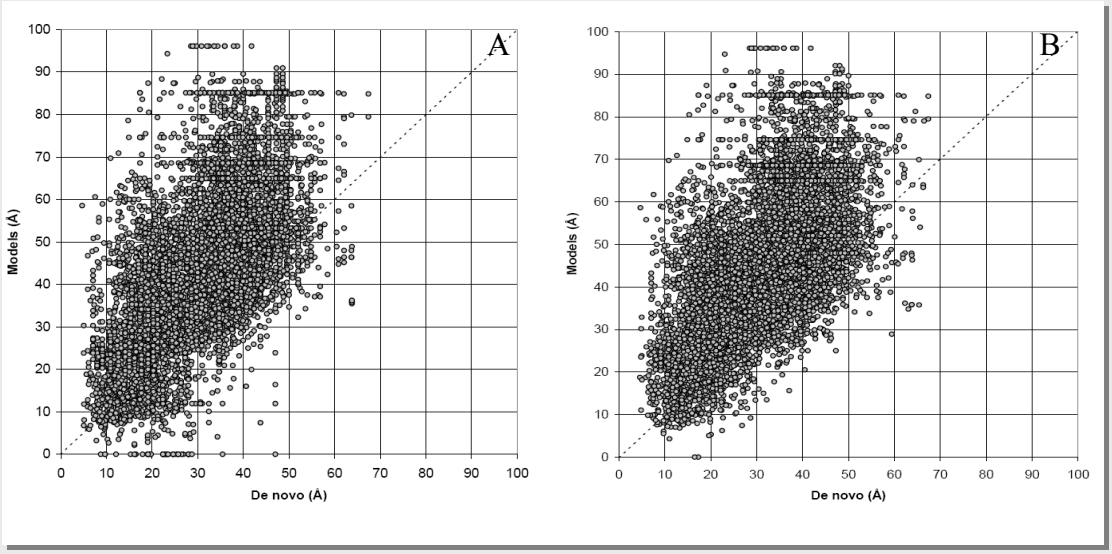
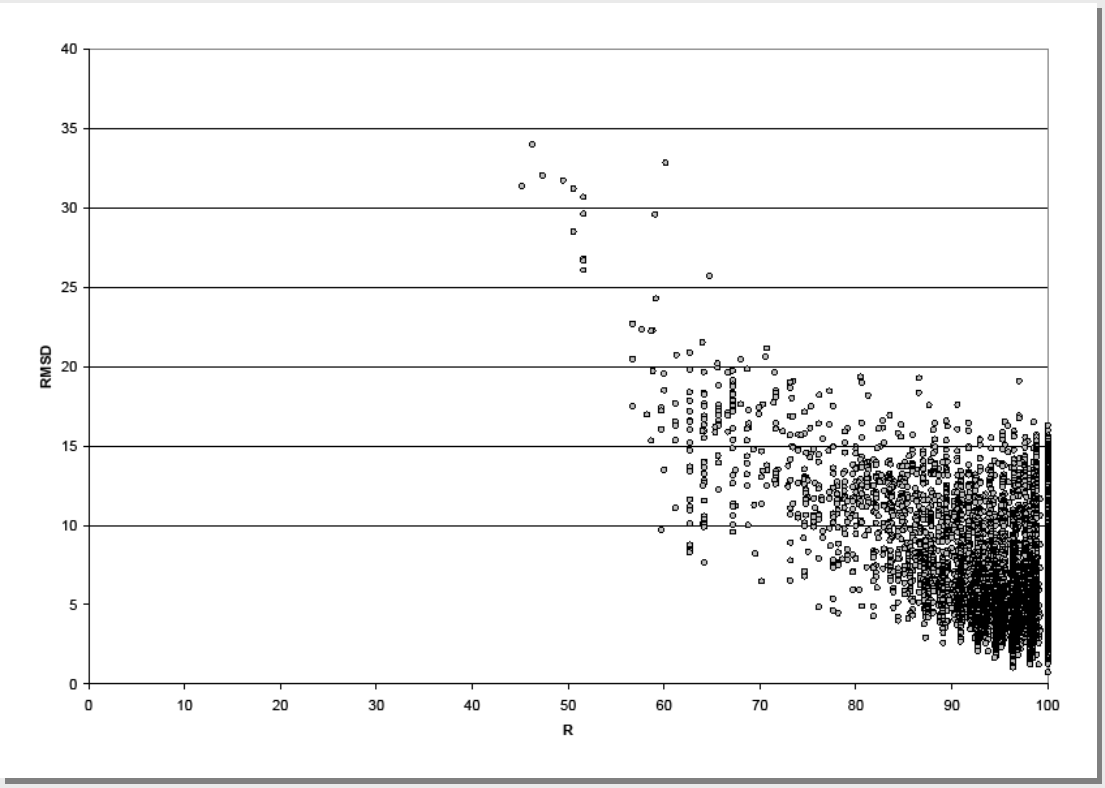


Figura 5.6. Existeix una relació inversa entre RMSD i fracció de residus alineats, això explica els valors de RMSD elevats mostrats a la figura 5.4A.



5.4. DISCUSSIÓ

L´estudi realitzat en aquest capítol complementa l´estudi del capítol anterior, i mostra la utilitat dels mètodes de comparació estructural alhora de detectar les parts de bona qualitat en prediccions *de novo*, i el seu posterior ús en el refinament de l´estructura.

En aquest capítol presentem un mètode de refinament basat en l'ús d´alineaments estructurals i modelat per homologia. El principal avantatge del mètode de refinament és la seva senzillesa a nivell conceptual i tècnic: únicament requereix la construcció de models per homologia utilitzant alineaments estructurals generats per comparació entre la predicció *de novo* i el pertinent homòleg.

Els desavantatges són essencialment aquelles assumpcions fetes alhora de triar els homòlegs de les prediccions (mirar *Capítol 4).* De forma més concreta, una altra limitació del present estudi és que està realitzat únicament sobre 15 proteïnes. No obstant això, tenint en compte que en un futur no molt llunyà les assumpcions comentades segurament trobaran solució (mirar *Capítol 4*) el protocol descrit es pot presentar com una alternativa, però sobretot com un mètode complementari als mètodes de refinat actuals, ja que podria ser utilitzat en una primera fase de, per tal d´obtenir prediccions refinades de forma global, i en una segona etapa emprar les tècniques actuals per aconseguir un refinat més fi.

5.5. REFERÈNCIES

1. Bradley, P., et al., *Rosetta predictions in CASP5: successes, failures, and prospects for complete automation.* Proteins, 2003. **53 Suppl 6**: p. 457-68.
2. Skolnick, J., et al., *TOUCHSTONE: a unified approach to protein structure prediction.* Proteins, 2003. **53 Suppl 6**: p. 469-79.
3. Jones, D.T. and L.J. McGuffin, *Assembling novel protein folds from super-secondary structural fragments.* Proteins, 2003. **53 Suppl 6**: p. 480-5.
4. Fang, Q. and D. Shortle, *Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments.* Proteins, 2003. **53 Suppl 6**: p. 486-90.
5. Lee, M.R., et al., *Molecular dynamics in the endgame of protein structure prediction.* J Mol Biol, 2001. **313**(2): p. 417-30.
6. Xiang, Z., *Advances in homology protein structure modeling.* Curr Protein Pept Sci, 2006. **7**(3): p. 217-27.
7. Eyrich, V.A., D.M. Standley, and R.A. Friesner, *Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set.* J Mol Biol, 1999. **288**(4): p. 725-42.
8. Ortiz, A.R., A. Kolinski, and J. Skolnick, *Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments.* J Mol Biol, 1998. **277**(2): p. 419-48.
9. Monge, A., et al., *Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models.* J Mol Biol, 1995. **247**(5): p. 995-1012.
10. Lee, M.R., D. Baker, and P.A. Kollman, *2.1 and 1.8 A average C(alpha) RMSD structure predictions on two small proteins, HP-36 and s15.* J Am Chem Soc, 2001. **123**(6): p. 1040-6.
11. Misura, K.M. and D. Baker, *Progress and challenges in high-resolution refinement of protein structure models.* Proteins, 2005. **59**(1): p. 15-29.

12. Tsai, J., et al., *An improved protein decoy set for testing energy functions for protein structure prediction.* Proteins, 2003. **53**(1): p. 76-87.
13. Fiser, A., R.K. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Sci, 2000. **9**(9): p. 1753-73.
14. Rapp, C.S. and R.A. Friesner, *Prediction of loop geometries using a generalized born model of solvation effects.* Proteins, 1999. **35**(2): p. 173-83.
15. Xiang, Z., C.S. Soto, and B. Honig, *Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7432-7.
16. Zheng, Q. and D.J. Kyle, *Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings.* Proteins, 1996. **24**(2): p. 209-17.
17. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins, 2004. **55**(2): p. 351-67.
18. Li, W., Z. Liu, and L. Lai, *Protein loops on structurally similar scaffolds: database and conformational analysis.* Biopolymers, 1999. **49**(6): p. 481-95.
19. Wojcik, J., J.P. Mornon, and J. Chomilier, *New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification.* J Mol Biol, 1999. **289**(5): p. 1469-90.
20. Fidelis, K., et al., *Comparison of systematic search and database methods for constructing segments of protein structure.* Protein Eng, 1994. **7**(8): p. 953-60.
21. van Vlijmen, H.W. and M. Karplus, *PDB-based protein loop prediction: parameters for selection and methods for optimization.* J Mol Biol, 1997. **267**(4): p. 975-1001.
22. Ponder, J.W. and F.M. Richards, *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J Mol Biol, 1987. **193**(4): p. 775-91.
23. Fan, H. and A.E. Mark, *Refinement of homology-based protein structures by molecular dynamics simulation techniques.* Protein Sci, 2004. **13**(1): p. 211-20.

24. Cozzetto, D., et al., *Assessment of predictions in the model quality assessment category.* Proteins, 2007. **69 Suppl 8**: p. 175-83.

25. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

26. Pearl, F.M., et al., *The CATH database: an extended protein family resource for structural and functional genomics.* Nucleic Acids Res, 2003. **31**(1): p. 452-5.

27. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.

28. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.

29. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallographica, 1976. **32**: p. 922-923.

30. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.

CAPÍTOL 6.

EFECTE DE LES MUTACIONS SOBRE LES CAVITATS DEL LISOZIM HUMÀ

6.1. INTRODUCCIÓ

Com s´ha comentat anteriorment, un dels punts més destacables dels mètodes de predicció estructural és que permeten estudiar aspectes funcionals de les proteïnes en casos on no hi ha dades experimentals disponibles, o bé la seva obtenció és complicada. Per posar un exemple, tal i com ja s´ha comentat en el primer capítol, les prediccions estructurals, i en particular els models per homologia, s´utilitzen habitualment en la interpretació d´experiments amb mutants [1, 2], en el disseny racional de fàrmacs [1, 3] o en l´estudi de la base molecular de mutacions patològiques [4]. En aquestes i altres aplicacions, les conclusions extretes dependran de la qualitat de les prediccions generades; és aquí on rau la importància de disposar de bons mètodes que permetin avaluar-ne la qualitat.

Un dels aspectes més destacats de l´estructura de les proteïnes és la seva superfície, i dins la superfície ho són les cavitats. S´ha observat que dites cavitats, particularment les de major mida, tenen un clar valor funcional [5, 6]. Per posar un exemple, estudis realitzats en els darrers anys mostren que la cavitat més gran tendeix a ser la que inclou el centre funcional. Addicionalment, les cavitats menors poden albergar centres reguladors de la funció proteica per mitjà de mecanismes al.lostèrics [5, 7]. Des d´un punt de vista aplicat les cavitats de les proteïnes tenen un interès particular en el cas de disseny de fàrmacs, ja que el seu coneixement pot guiar-ne el propi disseny o modificació.

Aquests aspectes són els que donen a les cavitats un interès especial, en front de la proteïna en la seva totalitat, i són els que ens han portat a estudiar la resposta d´aquestes cavitats a les mutacions, tal i com es veurà en aquest capítol.

CAVITATS DE PROTEÏNES I MODELS EVOLUTIUS

Per tal de proporcionar una base sòlida per als models evolutius de les proteïnes és necessari tenir un coneixement a nivell molecular detallat sobre l´efecte de les mutacions. De fet, recentment s´ha realitzat un important esforç per introduir elements de biofísica que relacionin el paper de les mutacions amb processos evolutius [8-11]. D´aquesta manera, per exemple s´ha pogut estudiar l´efecte de les mutacions en l'eficàcia biològica [8-10], les propietats de les mutacions compensadores [9], etc.

Tot i que per descriure acuradament l´efecte de les mutacions sobre les proteïnes i la seva funció des d´un punt de vista biofísic es necessiten diverses variables, en molts casos és suficient emprar la $\Delta\Delta G$ (diferència d´energia lliure o estabilitat, entre una estructura nativa i una mutant). Això es basa en el suposat que existeix una relació entre $\Delta\Delta G$ i variació funcionals, segons la qual si puja la primera es veu afectada la segona [12, 13]. Existeixen no obstant, dades que suggereixen que aquesta correlació no és tant clara; per exemple, si comparem els valors de $\Delta\Delta G$ *in vitro* i activitat de lisozims mutants de fag T4 [14], veiem que no existeix cap correlació entre $\Delta\Delta G$ i activitat enzimàtica. Aquesta manca de tendència es posa de manifest també en estudis realitzats per Poteete i col.laboradors [15], on s´observa que mutacions desestabilitzadores en el lisozim de fag T4 (ex. Ile3Gly, amb $\Delta\Delta G$ = -2.1 kcal/mol) no cursen amb pèrdua d´activitat enzimàtica *in vivo*.

Totes aquestes dades evidencien que basar els models evolutius únicament en els valors de $\Delta\Delta G$ és insuficient per obtenir una visió completa de l´impacte de les mutacions sobre la funció de les proteïnes, i que es requereixen altres paràmetres relacionats amb la funció, estructura o mecanismes al.lostèrics [16-18]. En aquest treball enfoquem aquest tema estudiant l´efecte de les mutacions sobre una de les característiques estructurals funcionals més importants: les cavitats. Tal i com s´ha comentat, les cavitats juguen un paper fonamental en la funció de la proteïna, ja sigui

a nivell enzimàtic o a nivell d´interacció amb l´entorn (medi o altres proteïnes o macromolècules,...).

Existeix un gran número d´evidències experimentals que mostren, implícitament o explícita, que les propietats estructurals/funcionals de les cavitats poden ser fàcilment modificades per una mutació puntual. Algunes de les dades sobre aquest tema deriven d´estudis de disseny de proteïnes amb noves funcions [19, 20]; altres venen d´estudis biomèdics sobre l´efecte de mutacions sobre l´activitat de determinats enzims [16, 17, 21] o fins i tot d´estudis farmacogenètics que demostren que petits canvis en els centres actius de certs enzims es tradueixen en sensibilitats a fàrmacs diferents [22].

Com hem comentat, en aquest estudi ens ocupem de l´efecte de les mutacions en les cavitats de les proteïnes. Aquest tema ha estat abordat en nombrosos estudis, tant de tipus bàsic [23-26] com aplicat [19, 20, 27]. No obstant això, aquests estudis es centren en un número reduït de cavitats [19, 20, 27], i en cavitats internes en comptes d´externes [23-25], etc. En el nostre estudi hem intentat ampliar aquestes dades, i analitzar la resposta del patró de cavitats del lisozim a les mutacions. Les raons d´escollir el lisozim com a model han estat les següents:

- Les estructures de la forma nativa i d´un bon número de mutants està disponible gràcies al treball del grup de Yutani [28].
- Existeixen dues versions experimentals de l´estructura nativa: 1lz1 i 1rex [28, 29]. La resolució d´ambdues és alta, i ens poden servir per de control, per separar aquells efectes de les mutacions dels que podrien ser fruit d´origen tècnic.
- El lisozim humà és una proteïna d´un únic domini, per tant les cavitats que trobem no poden ser fruit de la separació de dos dominis.
- Els 116 mutants disponibles cobreixen diverses localitzacions estructurals, anant des de residus enterrats a completament accessibles, residus amb estructura secundària,...

- L´ús del lisozim humà permet repartir el conjunt de mutacions en neutrals i deletèries emprant mètodes d´anotació automàtics parametritzats amb dades humanes (veure *Metodologia*).

En els propers apartats es passa a detallar la metodologia i resultats obtinguts.

6.2. METODOLOGIA

MUTANTS DEL LISOZIM HUMÀ

En l´estudi hem comparat les cavitats de l´estructura nativa per al lisozim humà, codi PDB: 1lz1 [29] amb un conjunt de 116 mutants produïts pel grup de Yutani [30-46]. La majoria d´aquests mutants presenten mutacions simples, tot i que alguns casos són dobles/triples, i altres casos delecions/insercions.

Les estructures han estat recuperades de la base de dades PDB [47]. És important remarcar que no s´ha utilitzat cap model teòric.

CÀLCUL DE CAVITATS

El càlcul de cavitats s´ha realitzat amb el programa Surfnet [5]. Per una donada proteïna Surfnet proporciona un conjunt de cavitats, definides per un seguit d´àtoms. Cal dir però, que tractant-se el càlcul de cavitats d´un mètode numèric, no és totalment invariant a les rotacions de la proteïna; és a dir, si es rota la proteïna les cavitats poden variar lleugerament.

Per tal de corregir aquest petit efecte s´ha seguit un protocol de refinament senzill per cadascuna de les proteïnes per les quals s´havia de fer el càlcul de cavitats:

1. Generació de 25 rotàmers a l´atzar.
2. *Clusterització* de les cavitats per als 25 rotàmers emprant un algorisme UPGMA, amb un punt de tall de solapament de 0.7.
3. Per cada *cluster* de cavitats s´ha obtingut una llista d`àtoms corresponent a la unió de tots els àtoms que el conformen.

La inspecció visual de les cavitats resultants mostra que aquestes són correctes i no presenten resultats artefactuals. A la figura 6.2 es poden veure alguns exemples. Cal

remarcar que aquest protocol ha estat optimitzat per al cas que ens ocupa, per tant podria no funcionar en altres casos.

De les cavitats resultants per proteïna després d´haver aplicat el protocol de refinament s´han seleccionat les 10 més grans. D´aquesta manera s´exclouen aquelles cavitats més sensibles a les fluctuacions estructurals.

A la següent taula es presenta una llista de les cavitats del lisozim natiu obtingudes després d´aplicar el protocol de refinament i el filtrat per mida:

Taula 6.1. Llista de les cavitats utilitzades, amb el número d´àtoms que les conformen

Àtoms per cavitat									
C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
202	126	112	88	76	70	69	59	55	54

COMPARACIÓ DE LES CAVITATS

Per poder comparar les cavitats del lisozim amb les del mutant és necessari primer decidir quines són les cavitats equivalents entre les dues estructures. Els passos seguits per definir les equivalències entre cavitats són els següents:

- Comparació d´una cavitat X de la nativa amb totes les cavitats d´un model Y.
- Per cada comparació es calcula el número d´àtoms en comú entre les dues cavitats.
- Aquelles comparacions amb un número d`àtoms en comú més alt donaran un parell de cavitats equivalents.

Aquests passos es repeteixen per cadascuna de les cavitats del lisozim, amb cadascun dels 116 mutants.

L´estructura del lisozim pot presentar certs àtoms com $O^{\delta 1}$ o $N^{\delta}2$ (residus d´asparagina) l´assignació dels quals és arbitrària. Per tal d´evitar aquest problema, aquest àtoms han estat eliminats de les cavitats del lisozim. De totes maneres el número d´àtoms eliminats és reduït, i no afecta a les conclusions finals.

Una vegada obtinguda la llista de cavitats equivalents entre el lisozim i els mutants, les hem comparat utilitzant dues mesures: solapament (àtoms en comú) i RMSD.

Solapament: número d´àtoms en comú entre una cavitat a la proteïna nativa i la cavitat equivalent al mutant dividida per la mida de la cavitat nativa.

RMSD: el càlcul d´RMSD s´ha obtingut superposant els àtoms comuns entre la cavitat nativa i la cavitat mutant equivalent.

Per tal de separar l´efecte de les mutacions del possible efecte d´origen tècnic (condicions experimentals diferents als mutants, per exemple) hem utilitzat com a valors de referència els resultats de comparar dues formes natives del lisozim: 1lz1 [29] i 1rex [48]. Així, per una determinada cavitat, el solapament i RMSD es poden expressar com:

$$Solapament_{1lz1-mutant} - Solapament_{1lzi-1rex}$$

$$RMSD_{1lz1-mutant} - RMSD_{1lzi-1rex}$$

Per un mutant determinat, valors positius en solapament i negatius per RMSD indiquen que l´efecte de la mutació és tant petit que no pot ser distingit d´efectes d´origen tècnics.

FLUCTUACIONS TÈRMIQUES ATÒMIQUES

Per tal d´obtenir una idea aproximada de les fluctuacions tèrmiques dels àtoms es va emprar la següent fórmula [49]: Fluctuació = $B/(8.\pi^2)$, on B correspon al factor de temperatura dels àtoms que podem trobar als arxius PDB.

FREQÜÈNCIA D´INTEREACCIONS MODIFICADES AL VOLTANT DEL PUNT DE MUTACIÓ

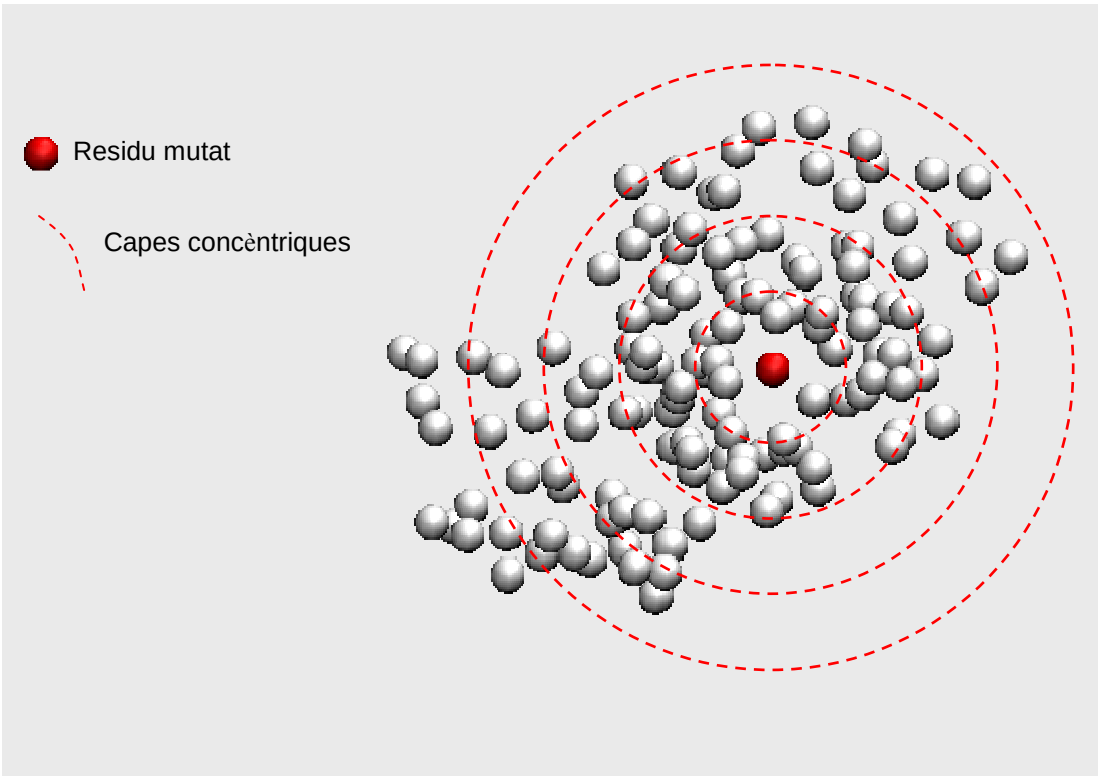
Aquest càlcul permet aportar evidències addicionals de l´efecte essencialment local de les mutacions a nivell d´interaccions residu-residu. Aquestes freqüències s´han mesurat per mitjà del potencial empíric ProsaII [50-52], només per als mutants simples (per evitar solapaments i desviacions en la conservació estructural).

S´ha seguit el següent protocol:

1. Càlcul de potencial ProsaII per a les dues versions de lisozim (1lz1 i 1rex).
2. Càlcul de potencial ProsaII per a 97 mutants (conjunt que presenta les mutacions simples).
3. Per cadascun dels mutants s´ha calculat la diferència d´energia dels seus residus respecte 1lz1: $\Delta e(i) = | e_{mut}(i) - e_{1lz1}(i)|$, on i és el número de residu. S´ha descartat la diferència quan el residu *i* correspon al residu mutat.
4. Aquests valors de Δe s´han classificat en capes, segons la distància del residu corresponent respecte el residu mutat.
5. Dins de cada capa es descarten els valors d´energia que estan per sota d´un llindar determinat.
6. Els valors que queden per cada capa es normalitzen pel volum seguint el procediment de Taylor i Kennard [53]. El resultat final és una freqüència d´energies per sobre un llindar per cada capa.

El protocol seguit es resumeix en el següent esquema:

Figura 6.1. Protocol seguit per al càlcul de la freqüència d´interaccions modificades al voltant del punt de mutació. A cada capa es mira Δe(i), i es compten els valors que estan per sobre d´un llindar obtingut comparant les dues formes natives (1rex i 1lz1).

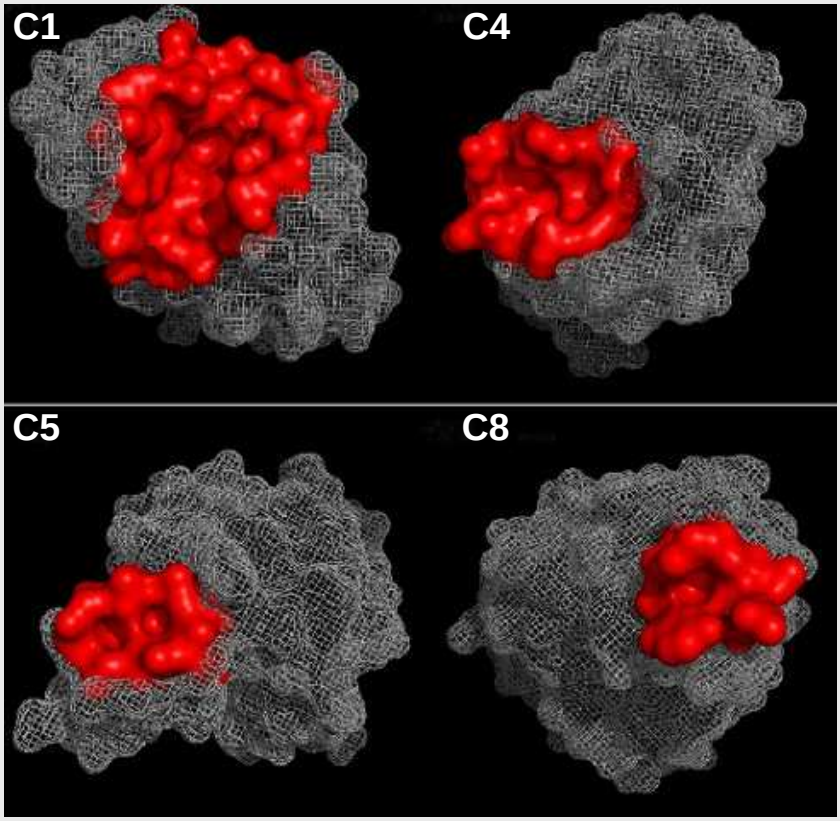


El llindar comentat al pas 5 s´utilitza per descartar variacions d´energia petites que podrien ser d´origen tècnic i no biològic (ex diferències en el procés de determinació d´estructura). El llindar és diferent per cada capa, i es calcula en base les diferències d´energia entre les dues formes natives 1rex i 1lz1. El llindar correspon al valor $\Delta e(i)_{1rex-1lzi}$ màxim de la capa, considerant el màxim tal i com es calcula en un *boxplot* (mirar *Metodologia general*).

6.3. RESULTATS

Com s´ha esmentat a la *Metodologia* hem treballat sobre les 10 cavitats més grans del lisozim humà (Taula 6.1); la resta no han estat considerades ja que poden ser més sensibles a fluctuacions estructurals. A la figura 6.2 es mostren 4 de les cavitats estudiades.

Figura 6.2. Exemples de cavitats del lisozim humà. Es mostren 4 de les cavitats emprades en l´estudi. Els àtoms de les cavitats es mostren en vermell.



Tal i com s´ha comentat al corresponent apartat de la *Metodologia*, s´han comparat les cavitats del lisozim amb aquelles equivalents en els mutants emprant dos paràmetres: solapament i RMSD, corregits amb els corresponents valors obtinguts per 1rex.

En els següents apartats es passa a descriure els efectes generals de les mutacions sobre les cavitats del lisozim, així com la relació d´aquests canvis amb factors com la localització de la mutació o tipus de mutació (deletèria o neutra).

EFECTES GENERALS DE LES MUTACIONS

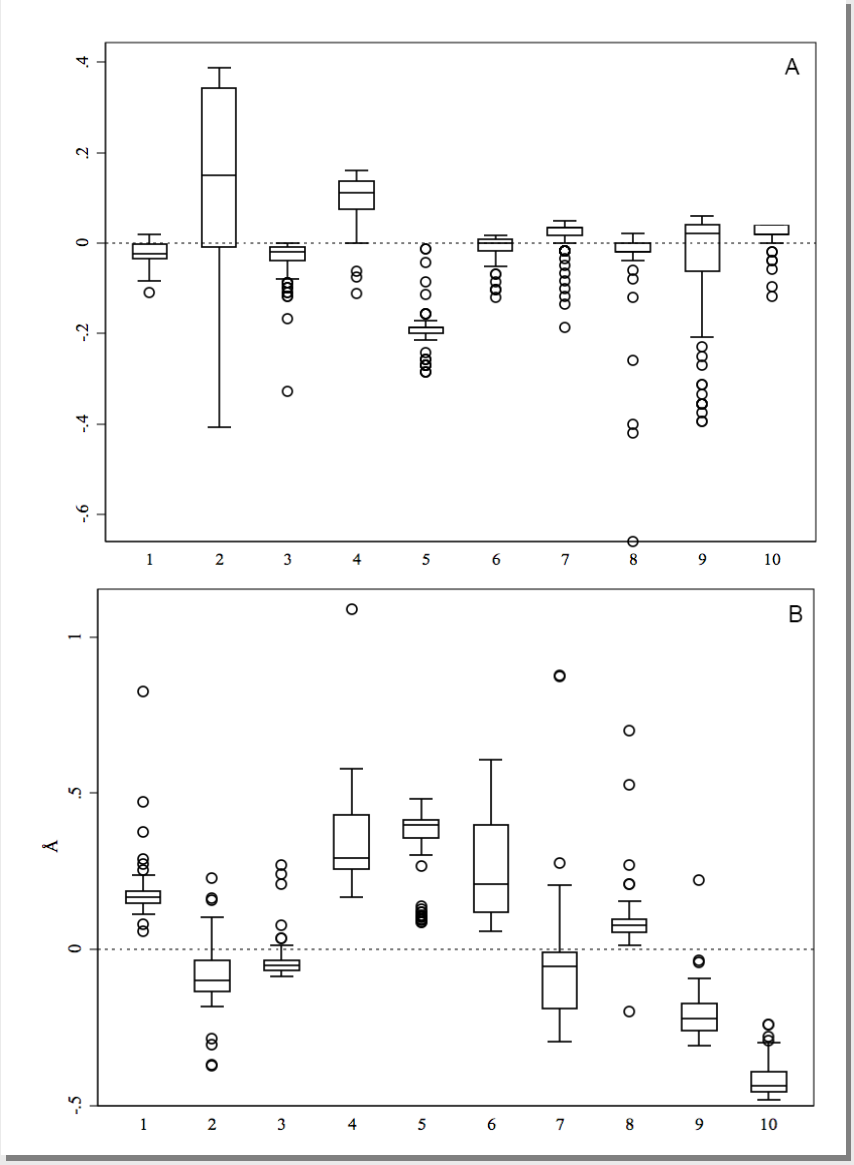
Trobem dues situacions similars tant pel solapament com per l´RMSD: figures 6.3A i 6.3B respectivament.

S´observa que tant les distribucions de solapament com les d´RMSD presenten molts punts que es desvien de la referència (línia discontinua); aquest fet ens indica que per la majoria de cavitats hi ha mutants amb la capacitat de canviar-ne la composició i/o forma. En general aquests casos corresponen a aquells mutants amb delecions o insercions; això confirma que aquest tipus de mutacions introdueixen canvis més dràstics a nivell estructural, que per tant podrien tenir repercussions a nivell funcional [54].

De totes maneres no totes les cavitats són igual de sensibles a les mutacions; de fet només unes quantes són afectades per totes les mutacions i presenten distribucions allunyades de la referència. Per a la seva discussió ens centrarem en aquelles cavitats amb Q3 (tercer quartil) més baix que la referència en el cas del solapament, i cavitats amb Q1 (primer quartil) més alt que la referència en el cas de l´RMSD. En base a aquests criteris podem dir que les cavitats C1, C3, C5 i C8 en el cas de solapament, i cavitats C1, C4, C5, C6 i C8 en el cas de l´RMSD són més sensibles a les mutacions que la resta. Més concretament, de totes aquestes, la C1, C3 i C5 són les més sensibles a l´efecte de les mutacions, ja que experimenten canvis importants a nivell de solapament i geometria. Altres cavitats a part d´aquestes també poden ser sensitives a les mutacions; tot i que probablement l´efecte de la mutació sigui més depenent de la localització del residu alterat. La cavitat 5 mereix una atenció especial, ja que és la que presenta unes variacions respecte la referència més importants,

sobretot de solapament. Aquest fet segurament es pot atribuir al fet que aquesta cavitat es troba a una zona de la proteïna amb poques restriccions estructurals.

Figura 6.3. Distribucions per al solapament i RMSD. Per les 10 cavitats del lisozim (C1 a C10) es mostren els resultants de solapament (A) i RMSD (B). Valors positius per solapament i negatius per RMSD indiquen que l´efecte de la mutació no pot ser distingit d´un possible efecte d´orígen tècnic.



En general, fins i tot per les cavitats més sensibles a mutacions els efectes observats són petits; això és consistent amb el fet que l´estructura global dels mutants és altament conservada en tots els casos. En el cas de l´RMSD els canvis de les cavitats van bàsicament de 0.2Å a 0.4Å; aquest rang és del mateix ordre que la mitjana de fluctuacions tèrmiques atòmiques observades per la proteïna nativa 1lz1, 0.5±0.2Å (veure *Metodologia*). Per tant no podem pensar que aquests canvis puguin tenir un efecte substancial sobre propietats geomètriques de les cavitats.

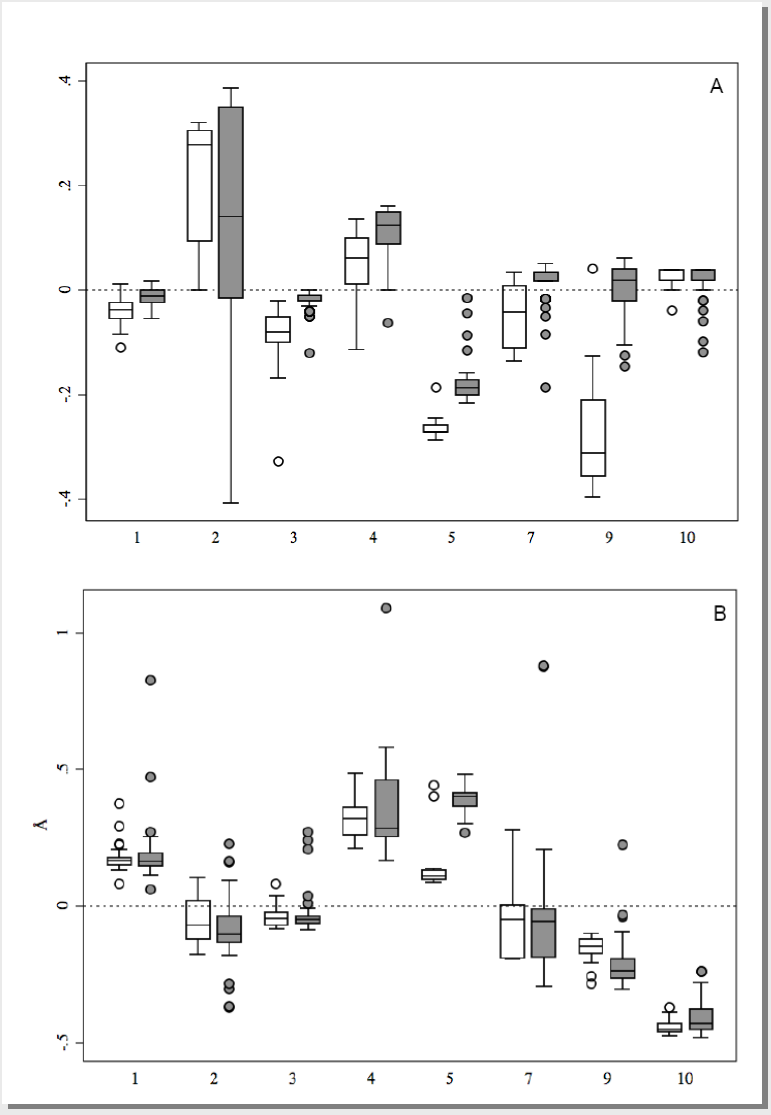
Si considerem el solapament la situació és diferent, ja que per molt petits que siguin els canvis de solapament ja impliquen canvis composicionals; i això junt amb el canvi ocasionat per la mutació pot alterar les propietats físico-químiques de la cavitat. Cal remarcar però que aquest punt pot variar segons les característiques de la mutació, com es tractarà tot seguit.

EFECTE DE LA LOCALITZACIÓ DE LA MUTACIÓ

Diversos estudis d´enginyeria de proteïnes han mostrat que els efectes de les mutacions poden ser o bé locals [30, 31, 55-57] o bé de llarg abast [55-59]. Per tal de determinar ambdues possibilitats en el cas del lisozim humà hem repartit les dades segons si la mutació cau dins o fora de cada cavitat. A les figures 6.4A i 6.4B es poden veure els resultats per a solapament i RMSD, respectivament.

S´observa una diferència clara entre solapament i RMSD. Quan les mutacions tenen lloc dins la cavitat, els corresponents valors de solapament s´allunyen més de la referència (figura 6.4A). Contràriament, no hi ha una diferència clara en el cas de l´RMSD, i tant si la mutació és dins com fora de la cavitat els valors són similars (figura 6.4B).

Figura 6.4. Dependència del solapament i RMSD respecte la localització de la mutació. Es comparen els valors de solapament (A) i RMSD (B) per les mutacions que tenen lloc dins la cavitat (caixes blanques) i mutacions que tenen lloc fora (caixes grises). Les cavitats 6 i 8 no s´han representat ja que tenien menys de 5 mutants amb residus dins les seves cavitats.

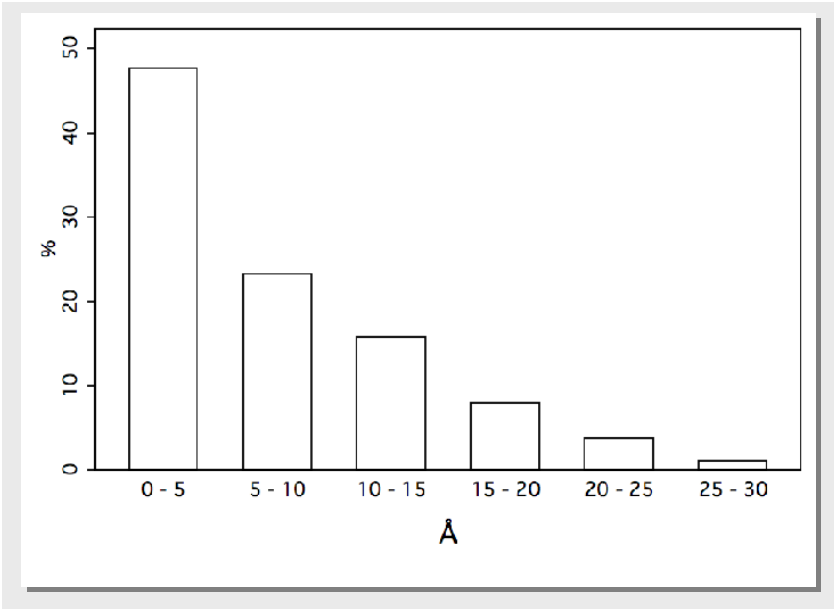


Mirats de forma global, els nostres resultats indiquen que tot i que les cavitats poden experimentar canvis per mutacions en punts allunyats, com descriuen Sinha i Nussinov [59], és més habitual que la causa dels canvis a les cavitats siguin les mutacions que les afecten directament. A més, aquests canvis probablement siguin molt locals, i estiguin principalment relacionats a canvis en les propietats de les cadenes laterals

canviades per la mutació: per exemple variacions de volum associades a canvi de cadenes laterals de diferent mida o variacions en la hidrofobicitat associades a canvi de cadenes alifàtiques per polars.

Per tal d´aportar més evidències sobre l´efecte principalment local de les mutacions hem emprat un potencial estadístic (veure *Metodologia*). Amb aquest potencial hem calculat la freqüència amb la qual les interaccions residu-residu canvien entre mutant i estructura nativa a mesura que ens allunyem del punt de la mutació. Com es pot veure a la figura 6.5, els canvis en las interaccions residu-residu són clarament més freqüents quan ens fixem en zones properes al punt de mutació. Això dóna validesa a la idea que en general els efectes principals de les mutacions puntuals són locals: a causa del canvi en les propietats de l´aminoàcid, i a causa de les modificacions ocasionades en la xarxa d´interaccions residu-residu.

Figura 6.5. Freqüència de canvis d´energia al voltant del punt de mutació. La figura mostra com els canvis d´energia es distribueixen a mesura que ens allunyem del punt de mutació. Les dades estan representades en capes de 5Å. Així per exemple en el rang 0-5 trobem les freqüències de canvi d´energia, per sobre un llindar concret, d´aquells residus que estan entre 0 i 5Å del residu mutat (distància Cα-Cα).



EFECTE DE LA NATURALESA DE LA MUTACIÓ

Les mutacions deletèries estan directament associades a dany sobre les proteïnes, ja sigui per modificació de la seva estabilitat o per alteració directa de les característiques estructurals [16-18, 21]. En aquesta secció comparem l´efecte de les mutacions deletèries i mutacions neutres sobre les cavitats del lisozim. S´ha dividit el conjunt de dades segons si la mutació és neutra o deletèria, segons dos criteris diferents:

ΔΔG

Canvis de $\Delta\Delta G \leq 0.25$ Kcal/mol s´associen a mutacions neutres; valors de $\Delta\Delta G > 0.25$ Kcal/mol s´associen a mutacions deletèries. El llindar de 0.25 Kcal/mol s´ha agafat de forma arbitrària, de tal manera que existeixi un compromís entre la necessitat de garantir un impacte menor de la mutació a l´estabilitat de lisozim, i un mínim conjunt de dades de cada tipus. Aquest criteri de classificació dóna 24 mutacions neutres i 72 mutacions deletèries (per un dels mutants no s´ha disposat de valors de $\Delta\Delta G$, per tant no ha estat utilitzat en aquest cas).

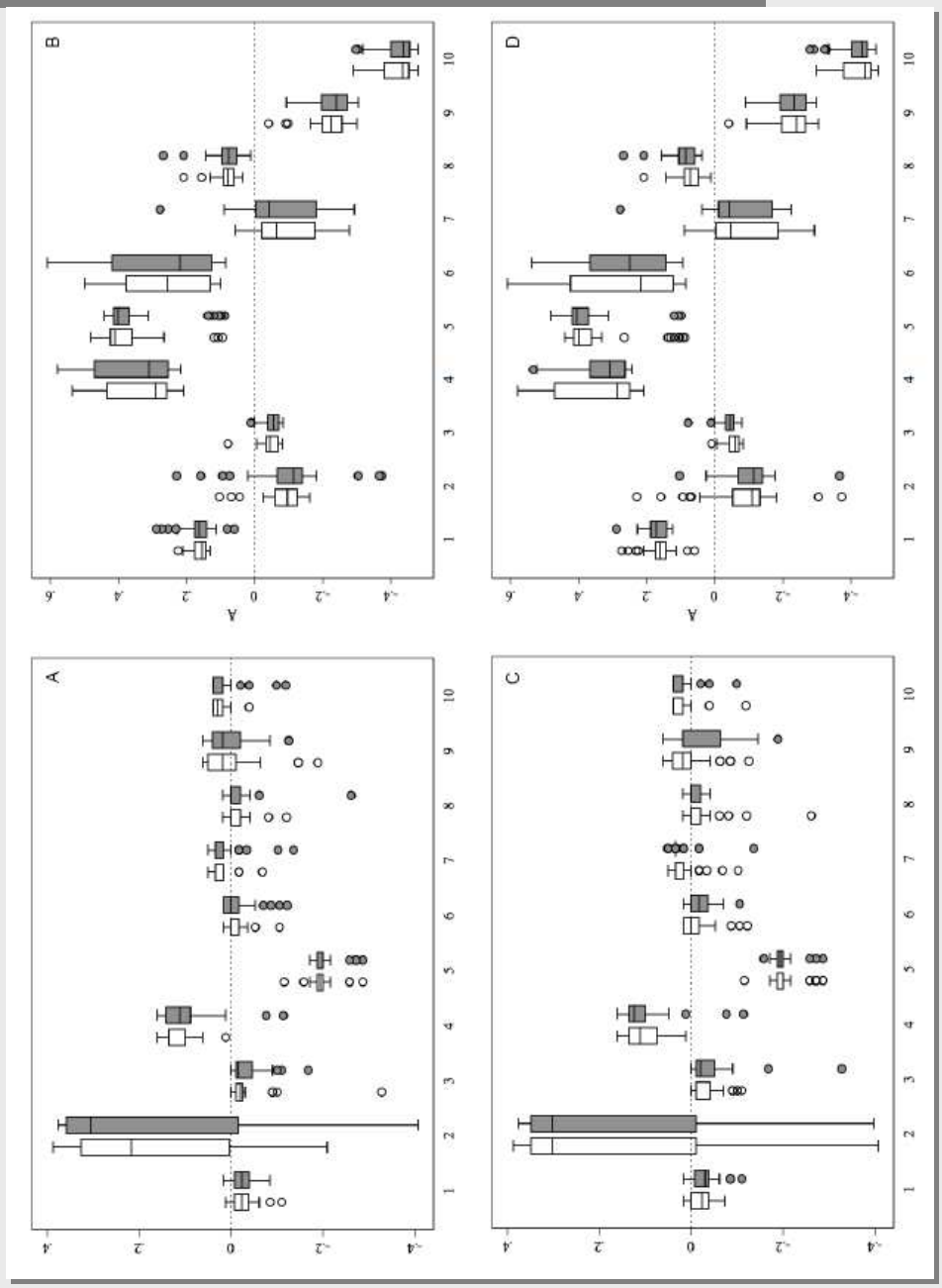
Notació automàtica

Aquest segon criteri està basat en un procediment de predicció desenvolupat al laboratori [60, 61]. Aquest mètode automàtic considera la conservació de propietats fisicoquímiques i la conservació d`aminoàcids entre espècies al punt de mutació, per discriminar entre mutacions associades a malalties i mutacions neutres [61]. Ha estat parametritzat emprant:

a) Mutacions associades a malalties Mendelianes [60, 61]. Corresponen a les mutacions deletèries.

b) Mutacions neutres corresponent a canvis observats entre proteïnes homòlogues properes (identitat de seqüència > 95%). Corresponen a les mutacions neutres.

Figura 6.6. Dependència del solapament i RMSD respecte la naturalesa de la mutació. S´han separat les mutacions en dos tipus: neutres (caixes blanques) i deletèries (caixes negres), emprant dos criteris: A i B corresponen a dades per solapament i RMSD respectivament, emprant com a criteri $\Delta\Delta G$; C i D corresponen a dades per solapament i RMSD respectivament, emprant com a criteri un sistema de notació automàtica.



Després d´aplicar aquest criteri de classificació hem obtingut 59 mutacions neutres i 38 mutacions deletèries (notar que en hi ha un total de 97 mutacions, una més que per al mètode basat en $\Delta\Delta G$, com s´ha comentat anteriorment).

Els dos models donen resultats similars (figura 6.6), amb poblacions per mutacions deletèries/neutres que mostren poques diferències entre elles, tant pel solapament com per l´RMSD. Si ens centrem en el solapament, pel qual hi havia clares diferències quan consideràvem l´efecte de la localització de la mutació (figura 6.4A), les diferències segons el tipus de mutació són inexistents (figures 6.6A per classificació segons $\Delta\Delta G$ i 6.6C per classificació segons notació automàtica). Si considerem específicament la partició de mutants en base a $\Delta\Delta G$, els nostres resultats constitueixen una extensió a nivell de cavitat de l´observació segons la qual no hi ha relació entre $\Delta\Delta G$ i els valors de RMSD globals dels mutants [62].

6.4. DISCUSSIÓ

Per tal de millorar el coneixement sobre l´evolució de les proteïnes, a part de fixar-nos en $\Delta\Delta G$ es requereixen nous paràmetres que permetin descriure l´efecte de les mutacions en la funcionalitat de la proteïna. La identificació d´aquests paràmetres passa per un enteniment sòlid i a nivell molecular de com les mutacions afecten a les característiques funcionals de les proteïnes, entre les quals les cavitats juguen un paper important. En aquest treball hem emprat el lisozim humà com a model sobre el qual estudiar l´efecte de les mutacions sobre les cavitats. Els resultats mostren que quan l´estructura del mutant es troba conservada (figura 6.3), les cavitats experimenten canvis lleus (corresponent els més importants a insercions/delecions). Existeix una certa dependència respecte la localització de la mutació, ja que quan aquesta es troba dins la mateixa cavitat els efectes són majors (figura 6.4), sobretot si ens fixem en la composició atòmica de la cavitat (figura 6.4A). Pel que fa a la naturalesa de la mutació (neutra o deletèria), no sembla jugar un paper important (figura 6.6) en els canvis de forma de la cavitat.

Així doncs podem concloure que les mutacions tenen un efecte principalment local (figura 6.5); addicionalment, els efectes que causen no estarien tant relacionats amb canvis geomètrics i composició, sinó que ho estarien amb canvis de propietats fisicoquímiques fruit del canvi de la cadena lateral del residu implicat.

El fet que les mutacions afectin més a la composició de les cavitats que a la seva forma, i sobretot a les propietats fisicoquímiques, suggereix que els models actuals sobre evolució de proteïnes haurien d´incorporar nous paràmetres per definir l´efecte de les mutacions. Així per exemple seria interessant tenir en compte:

- Localització de la mutació: permetria veure si la mutació afecta a zones de superfície.
- Paràmetres fisicoquímics: com per exemple variacions d´hidrofobicitat, propensió a estructura secundària, volum,...

Aquests paràmetres junt amb $\Delta\Delta G$ es podrien combinar per relacionar millor l´efecte de la mutació a nivell molecular amb el fenotip que causa la mateixa. De fet hi ha casos on s´ha dut a terme en mutacions associades a malalties [16, 18, 21, 61, 63].

Tot i que els resultats exposats a aquest capítol es basen en una única proteïna, podrien generalitzar-se a altres mutants on les estructures no experimenten alteracions massa importants [57, 63].

EVOLUCIÓ I AL.LOSTERISME

Una conseqüència addicional dels nostres resultats està relacionada amb l´evolució dels mecanismes al.lostèrics. En una publicació recent, Liang i col.laboradors [64] postulen que l´al.losterisme i la unió de lligand podrien haver aparegut de forma simultània a causa de l´efecte desestabilitzador de les mutacions. Segons aquests autors això és conseqüència de la capacitat de les mutacions de generar heterogenïcitat conformacional. No és del tot clar si aquesta teoria és generalitzable, ja que no existeix una relació òbvia entre canvis conformacionals i estabilitat proteica [57] com s´ha comentat a la introducció.

Pel que fa als nostres resultats, mostren que és relativament fàcil per una mutació modificar la composició d´una cavitat (Figura 6.4A), per tant les propietats fisicoquímiques al punt de mutació es veurien alterades. Això suggereix un altra possible situació per l´evolució de l´al.losterisme: l´evolució simplement utilitzaria l´acoblament preexistent entre cavitats, fruit de la naturalesa al.lostèrica de les proteïnes [65]; d´aquesta manera que si una cavitat secundària és modificada es

podria crear un punt d´unió a un nou substrat sense alterar cap propietat estructural (tal i com mostren els nostres resultats); la unió del nou substrat podria aleshores induir un canvi conformacional sobre la cavitat principal a través de l´acoblament existent entre ambdues, modificant l´activitat d´aquesta.

6.5. REFERÈNCIES

1. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
2. Gillece, P., et al., *Export of a cysteine-free misfolded secretory protein from the endoplasmic reticulum for degradation requires interaction with protein disulfide isomerase.* J Cell Biol, 1999. **147**(7): p. 1443-56.
3. Marti, L., et al., *Exploring the binding mode of semicarbazide-sensitive amine oxidase/VAP-1: identification of novel substrates with insulin-like activity.* J Med Chem, 2004. **47**(20): p. 4865-74.
4. Sali, A., *Modeling mutations and homologous proteins.* Curr Opin Biotechnol, 1995. **6**(4): p. 437-51.
5. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. **13**(5): p. 323-30, 307-8.
6. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Sci, 1998. **7**(9): p. 1884-97.
7. DesJarlais, R.L., et al., *Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure.* J Med Chem, 1988. **31**(4): p. 722-9.
8. DePristo, M.A., D.M. Weinreich, and D.L. Hartl, *Missense meanderings in sequence space: a biophysical view of protein evolution.* Nat Rev Genet, 2005. **6**(9): p. 678-87.
9. Ferrer-Costa, C., M. Orozco, and X. de la Cruz, *Characterization of compensated mutations in terms of structural and physico-chemical properties.* J Mol Biol, 2007. **365**(1): p. 249-56.
10. Counago, R., S. Chen, and Y. Shamoo, *In vivo molecular evolution reveals biophysical origins of organismal fitness.* Mol Cell, 2006. **22**(4): p. 441-9.

11. Bloom, J.D., et al., *Protein stability promotes evolvability.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5869-74.

12. Meiering, E.M., L. Serrano, and A.R. Fersht, *Effect of active site residues in barnase on activity and stability.* J Mol Biol, 1992. **225**(3): p. 585-9.

13. Shoichet, B.K., et al., *A relationship between protein stability and protein function.* Proc Natl Acad Sci U S A, 1995. **92**(2): p. 452-6.

14. Matthews, B.W., *Studies on protein stability with T4 lysozyme.* Adv Protein Chem, 1995. **46**: p. 249-78.

15. Rennell, D., et al., *Systematic mutation of bacteriophage T4 lysozyme.* J Mol Biol, 1991. **222**(1): p. 67-88.

16. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.

17. Ferrer-Costa, C., M. Orozco, and X. de la Cruz, *Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.* J Mol Biol, 2002. **315**(4): p. 771-86.

18. Schuster-Bockler, B. and A. Bateman, *Protein interactions in human genetic diseases.* Genome Biol, 2008. **9**(1): p. R9.

19. Roncone, R., et al., *Engineering peroxidase activity in myoglobin: the haem cavity structure and peroxide activation in the T67R/S92D mutant and its derivative reconstituted with protohaemin-l-histidine.* Biochem J, 2004. **377**(Pt 3): p. 717-24.

20. Swift, J., et al., *Design of functional ferritin-like proteins with hydrophobic cavities.* J Am Chem Soc, 2006. **128**(20): p. 6611-9.

21. Sunyaev, S., V. Ramensky, and P. Bork, *Towards a structural basis of human non-synonymous single nucleotide polymorphisms.* Trends Genet, 2000. **16**(5): p. 198-200.

22. Weinshilboum, R.M. and L. Wang, *Pharmacogenetics and pharmacogenomics: development, science, and translation.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 223-45.

23. Buckle, A.M., P. Cramer, and A.R. Fersht, *Structural and energetic responses to cavity-creating mutations in hydrophobic cores: observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities.* Biochemistry, 1996. **35**(14): p. 4298-305.

24. Quillin, M.L., et al., *Size versus polarizability in protein-ligand interactions: binding of noble gases within engineered cavities in phage T4 lysozyme.* J Mol Biol, 2000. **302**(4): p. 955-77.

25. Lee, J., K. Lee, and S. Shin, *Theoretical studies of the response of a protein structure to cavity-creating mutations.* Biophys J, 2000. **78**(4): p. 1665-71.

26. Machicado, C., M. Bueno, and J. Sancho, *Predicting the structure of protein cavities created by mutation.* Protein Eng, 2002. **15**(8): p. 669-75.

27. Maglio, O., et al., *Preorganization of molecular binding sites in designed diiron proteins.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3772-7.

28. Funahashi, J., et al., *How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme.* Protein Eng, 2003. **16**(9): p. 665-71.

29. Artymiuk, P.J. and C.C. Blake, *Refinement of human lysozyme at 1.5 A resolution analysis of non-bonded and hydrogen-bond interactions.* J Mol Biol, 1981. **152**(4): p. 737-62.

30. Takano, K., et al., *Contribution of hydrophobic residues to the stability of human lysozyme: calorimetric studies and X-ray structural analysis of the five isoleucine to valine mutants.* J Mol Biol, 1995. **254**(1): p. 62-76.

31. Yamagata, Y., et al., *Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six tyrosine --> phenylalanine mutants.* Biochemistry, 1998. **37**(26): p. 9355-62.

32. Funahashi, J., et al., *The structure, stability, and folding process of amyloidogenic mutant human lysozyme.* J Biochem, 1996. **120**(6): p. 1216-23.

33. Takano, K., et al., *Contribution of water molecules in the interior of a protein to the conformational stability.* J Mol Biol, 1997. **274**(1): p. 132-42.

34. Kuroki, R. and K. Yutani, *Structural and thermodynamic responses of mutations at a Ca2+ binding site engineered into human lysozyme.* J Biol Chem, 1998. **273**(51): p. 34310-5.

35. Takano, K., Y. Yamagata, and K. Yutani, *A general rule for the relationship between hydrophobic effect and conformational stability of a protein: stability and structure of a series of hydrophobic mutants of human lysozyme.* J Mol Biol, 1998. **280**(4): p. 749-61.

36. Takano, K., et al., *Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six Ser --> Ala mutants.* Biochemistry, 1999. **38**(20): p. 6623-9.

37. Takano, K., et al., *Effect of foreign N-terminal residues on the conformational stability of human lysozyme.* Eur J Biochem, 1999. **266**(2): p. 675-82.

38. Takano, K., et al., *Experimental verification of the 'stability profile of mutant protein' (SPMP) data using mutant human lysozymes.* Protein Eng, 1999. **12**(8): p. 663-72.

39. Funahashi, J., et al., *Role of surface hydrophobic residues in the conformational stability of human lysozyme at three different positions.* Biochemistry, 2000. **39**(47): p. 14448-56.

40. Takano, K., et al., *Contribution of salt bridges near the surface of a protein to the conformational stability.* Biochemistry, 2000. **39**(40): p. 12375-81.

41. Takano, K., Y. Yamagata, and K. Yutani, *Role of amino acid residues at turns in the conformational stability and folding of human lysozyme.* Biochemistry, 2000. **39**(29): p. 8655-65.

42. Goda, S., et al., *Effect of extra N-terminal residues on the stability and folding of human lysozyme expressed in Pichia pastoris.* Protein Eng, 2000. **13**(4): p. 299-307.

43. Takano, K., Y. Yamagata, and K. Yutani, *Role of amino acid residues in left-handed helical conformation for the conformational stability of a protein.* Proteins, 2001. **45**(3): p. 274-80.

44. Takano, K., Y. Yamagata, and K. Yutani, *Contribution of polar groups in the interior of a protein to the conformational stability.* Biochemistry, 2001. **40**(15): p. 4853-8.

45. Takano, K., Y. Yamagata, and K. Yutani, *Role of non-glycine residues in left-handed helical conformation for the conformational stability of human lysozyme.* Proteins, 2001. **44**(3): p. 233-43.

46. Funahashi, J., et al., *Positive contribution of hydration structure on the surface of human lysozyme to the conformational stability.* J Biol Chem, 2002. **277**(24): p. 21792-800.

47. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

48. Muraki, M., et al., *Origin of carbohydrate recognition specificity of human lysozyme revealed by affinity labeling.* Biochemistry, 1996. **35**(42): p. 13562-7.

49. Petsko, G.A. and D. Ringe, *Fluctuations in protein structure from X-ray diffraction.* Annu Rev Biophys Bioeng, 1984. **13**: p. 331-71.

50. Casari, G. and M.J. Sippl, *Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds.* J Mol Biol, 1992. **224**(3): p. 725-32.

51. Sippl, M.J., *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.* J Mol Biol, 1990. **213**(4): p. 859-83.

52. Sippl, M.J., et al., *Helmholtz free energies of atom pair interactions in proteins.* Fold Des, 1996. **1**(4): p. 289-98.

53. Taylor, R. and O. Kennard, *Hydrogen-bond geometry in organic-crystals.* Accounts Chem Res, 1984. **19**: p. 320-326.

54. Shortle, D. and J. Sondek, *The emerging role of insertions and deletions in protein engineering.* Curr Opin Biotechnol, 1995. **6**(4): p. 387-93.

55. Alber, T. and B.W. Matthews, *Structure and thermal stability of phage T4 lysozyme.* Methods Enzymol, 1987. **154**: p. 511-33.

56. Alber, T., *Mutational effects on protein stability.* Annu Rev Biochem, 1989. **58**: p. 765-98.

57. Shortle, D., *Mutational studies of protein structures and their stabilities.* Q Rev Biophys, 1992. **25**(2): p. 205-50.

58. Takano, K., et al., *Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants.* Biochemistry, 1997. **36**(4): p. 688-98.

59. Sinha, N. and R. Nussinov, *Point mutations and sequence variability in proteins: redistributions of preexisting populations.* Proc Natl Acad Sci U S A, 2001. **98**(6): p. 3139-44.

60. Ferrer-Costa, C., et al., *PMUT: a web-based tool for the annotation of pathological mutations on proteins.* Bioinformatics, 2005. **21**(14): p. 3176-8.

61. Ferrer-Costa, C., M. Orozco, and X. de la Cruz, *Sequence-based prediction of pathological mutations.* Proteins, 2004. **57**(4): p. 811-9.

62. Wallin, S., J. Farwer, and U. Bastolla, *Testing similarity measures with continuous and discrete protein models.* Proteins, 2003. **50**(1): p. 144-57.

63. Taverna, D.M. and R.A. Goldstein, *Why are proteins so robust to site mutations?* J Mol Biol, 2002. **315**(3): p. 479-84.

64. Liang, J., et al., *Ligand binding and allostery can emerge simultaneously.* Protein Sci, 2007. **16**(5): p. 929-37.

65. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins, 2004. **57**(3): p. 433-43.

CAPÍTOL 7.

CONSERVACIÓ DE LES CAVITATS EN ELS MODELS PER HOMOLOGIA

7.1. INTRODUCCIÓ

Tal i com s´ha comentat anteriorment, un dels reptes més importants de l´era post-genòmica és omplir l´enorme buit que existeix entre el gran número de seqüències conegudes, i el relativament reduït número d´estructures [1-4]. Des de la branca experimental s´han impulsat diversos projectes de genòmica estructural, que intenten abordar aquest repte per mitjà de processos massius de determinació estructural [5-11]. Els resultats d´aquests projectes ja són patents, ja que estudis recents demostren un increment considerable en el número d´estructures i particularment d´estructures que presenten nous *folds* o plegaments [12-16].

Tot i els grans avenços, i la important contribució no només d´aquests projectes de genòmica estructural, sinó també de grups experimentals individuals, proporcionar estructura a totes les proteïnes existents és ara per ara un objectiu impossible d´aconseguir. És per això que una de les vies utilitzades en aquests projectes de determinació massiva és l´ús d´eines de modelatge per homologia [1, 4-6, 9, 12, 17, 18]. Cal dir però que tot i que aquestes eines poden ser molt útils alhora de proporcionar prediccions estructurals, el seu rang d´aplicació pot variar substancialment [17, 19, 20].

En la següent taula es resumeixen les qualitats requerides per alguns dels camps on el modelatge per homologia és més utilitzada [5, 19-21].

Taula 7.1. Utilització dels models segons la qualitat

Finalitat	Qualitat requerida
Disseny de fàrmacs.	Alta; identitat entre *target* i *template* superior al 70%.
Interpretació de l´efecte de mutacions puntuals.	Mitjana/Alta; identitat entre *target* i *template* al voltant de 50%.
Disseny de pseudomolècules que s´uneixin al centre actiu d´un enzim.	Baixa; la identitat entre *target* i *template* al voltant de 30%.

La importància de saber la qualitat d´un model estructural és tal que en els darrers anys han sorgit nombrosos estudis [20, 22-25] destinats a calibrar la qualitat dels mètodes de predicció per homologia, i com s´ha comentat en el primer capítol fins i tot s´ha creat una categoria destinada a tal efecte als experiments CASP. De totes maneres la gran majoria d´aquests estudis es centren en la qualitat del model de forma global [17-20] i únicament uns pocs ho han fet en la qualitat d´aquelles parts de la proteïna que poden tenir rellevància biològica, com per exemple centres actius, cavitats o punts d´unió a altres molècules [22, 23, 26, 27].

Entre aquests resultats destaquen els obtinguts per De-Weese i Mout, que han aportat informació sobre com es preserva la qualitat de les cavitats en els models per homologia [28]. Per la seva part R. Sanchez i A. Sali, han aportat llum sobre alguns aspectes relacionats amb la conservació de cavitats en els models per homologia, centrant-se sobretot en la conservació dels contactes amb el substrat, que en el seu cas sol correspondre a una petita molècula orgànica. No obstant això aquest treball presenta algunes limitacions:

- Els resultats estan limitats pel reduït número de proteïnes i models estudiats (10 proteïnes i 207 models).
- La definició de cavitat és massa simple, ja que es considera una cavitat com un reduït conjunt de residus que presenten contactes amb petites molècules de lligands.
- L´anàlisi està essencialment basat en una variable: RMSD.

En el present capítol es descriu el treball realitzat amb l´objectiu d´ampliar aquests estudis i proporcionar una visió més detallada de l´efecte del modelat per homologia sobre les cavitats de les proteïnes. De forma més concreta el nostre objectiu és examinar la qualitat de les cavitats en models per homologia obtinguts a diferents rangs d´identitat –identitat de seqüència entre la proteïna a modelar i *template*

utilitzat–, emprant sis variables que cobreixen diverses característiques estructurals d´una cavitat.

Tot i que s´han obtingut dades per tot el rang d´identitats, tal i com es pot veure a la Taula 7.2, ens hem centrat en el comportament d´aquells models on la identitat de seqüència entre la proteïna a modelar i el motlle és mitjana (30%–60%) i baixa (<30%). Les raons són les següents:

- La qualitat dels models per homologia amb identitat de seqüència superior al 60% és habitualment elevada [17, 19].
- La funció bioquímica entre proteïnes amb identitats de seqüència per sobre el 60% tendeix a estar conservada [29, 30].
- La selecció de motlles en protocols de genòmica estructural solen tenir com a límit inferior el 30%, per tal d´assegurar una màxima cobertura [1, 31].
- L´estructura entre proteïnes amb identitats inferiors a 30% és conservada [29, 32, 33].

Taula 7.2. Distribució de rangs d´identitat de seqüència coverts.

IDENTITAT	FREQ. ABSOLUTA	FREQ. RELATIVA
0-10	12650	23.64
10-20	28382	53.04
20-30	4868	9.10
30-40	1830	3.42
40-50	1069	2.00
50-60	650	1.21
60-70	172	0.32
70-80	16	0.03
80-90	12	0.02
90-100	25	0.05
=100	3833	7.16

L´estudi s´ha dut a terme utilitzant un total de 53507 models per homologia construïts amb el programa Modeller [34]. Els resultats proporcionen una visió detallada i quantitativa de com la qualitat d´una cavitat varia amb la similitud que hi ha entre la nostre proteïna d´interès i l´homòleg d´estructura disponible, i constitueix una guia útil per als usuaris d´aquesta tècnica.

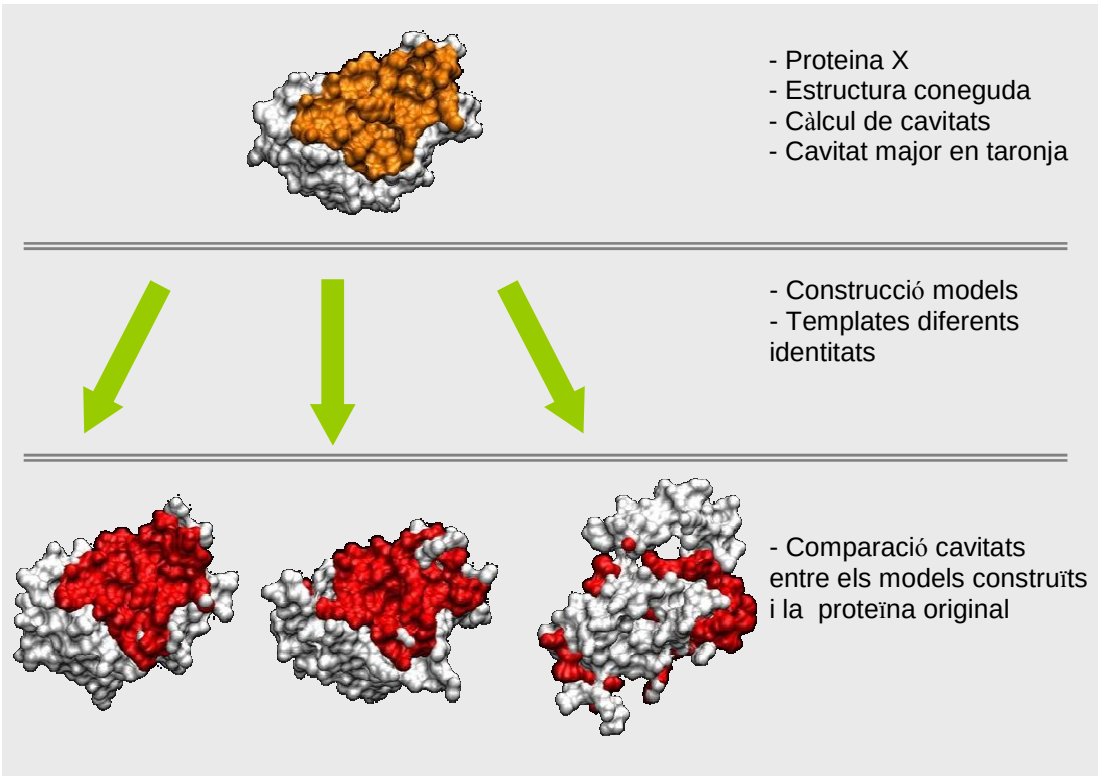
7.2. METODOLOGIA

Com s´ha comentat l´objectiu del treball és analitzar com varien les cavitats en un conjunt de models per homologia a mesura que varia la identitat de la proteïna modelada respecte el *template* que s´ha utilitzat.

SELECCIÓ DELS PARELLS TARGET-TEMPLATE

Els mètodes de modelatge per homologia es basen en la següent premissa: disposem d´una seqüència a modelar anomenada *target*, amb una certa identitat de seqüència respecte una altra proteïna amb estructura anomenada *template* (o motlle); si ambdues seqüències són similars (és a dir, la identitat de seqüència és elevada) podem pensar que les estructures també seran similars, per tant podem fer una transferència de l´estructura des del *template* cap al *target*.

Figura 7.1. Esquema del procés. Es calculen les cavitats sobre la proteina nativa; en una segona etapa es construeixen els models per homologia a diferents identitats de seqüència. Finalment es compara la forma de les cavitats natives amb les dels models.



Tenint en compte això, el primer que cal fer és construir una llista de parelles *target-templates* que cobreixin el rang d´identitats desitjat. Cal tenir present que en el nostre cas és indispensable disposar de l´estructura real de la proteïna *target* que volem modelar a diferents identitats de seqüència, ja que és fonamental per poder calibrar la qualitat dels models generats. El procés es troba resumit a la figura 7.1.

Per establir la llista original de proteïnes a modelar s´ha utilitzat la base de dades de dominis estructurals CATH v3.0.0 [35] (mirar *Metodologia general*). La llista de parelles *target-template* ha estat confeccionada agafant els dominis del nivell O (dominis que pertanyen a la mateixa topologia, i tenen una identitat de seqüència superior al 60%, veure *Metodologia general*) i agrupant-los segons el seu nivell H (Nivell de superfamília d´homologia). Aquests dominis de nivell O han estat obtinguts del mateix servidor de CATH, i han passat un seguit de filtres de qualitat:

- S´han exclòs aquells dominis discontinus.
- S´han exclòs aquells dominis considerats com a *folds* falsos a SCOP 1.67 [36].
- S´han exclòs aquells dominis amb aminoàcids incomplets, o discontinuïtats a la cadena principal.

El total de dominis després d´aplicar els filtres és de 3802; i després de fer les pertinents agrupacions segons el seu grup H tenim un total de 90948 parelles *target-template*. Aquests parells han estat utilitzats per construir els models utilitzant el programa Modeller, resultant en un total de 88410 models generats (la diferència es deu a certs alineaments que per presentar una diferència d´identitat massa elevada no han pogut ser efectuats pel programa).

Sobre els 88410 models s´ha aplicat un darrer filtre: només s´han conservat aquells models on tots els residus involucrats en la cavitat més gran estaven alineats. Aquest filtre és força selectiu, i el conjunt de models baixa fins a 53507. És sobre aquest darrer conjunt que s´han fet tots els càlculs.

PROTOCOL DE MODELAT PER HOMOLOGIA

Els models per homologia han estat obtinguts utilitzant el programa Modeller [34], emprant paràmetres per defecte. El procés de modelatge consta de dos passos: alineament entre el *target* i *template*, i transferència de les coordenades estructurals del *template* al *target* utilitzant com a pauta l´alineament generat.

La qualitat final del model depèn sobretot de la qualitat de l´alineament (d´aquí que una identitat elevada que asseguri un bon alineament donarà models de bona qualitat). Modeller [34] implementa un algorisme de programació global dinàmica amb penalitzacions per *gap* optimitzades per al modelat per homologia, ja que posiciona els *gaps* en un context estructural millor. Cal dir però que el rendiment dels algorismes de programació dinàmica com l´implementat per Modeller decreixen per identitats molt baixes (al voltant del 20%), i tot i que hi ha diverses alternatives per obtenir alineaments raonables a aquestes identitats [37] no és fàcil veure quines són millors. En el nostre cas es va optar, seguint estudis realitzats [23], per utilitzar alineaments basats en estructura per tal de mostrar com millora la qualitat d´un model si ho fa l´alineament, entenent que els alineaments estructurals són els millors alineaments que es poden obtenir. Concretament hem reconstruït tots els models però utilitzant un alineament estructural obtingut amb Mammoth [38, 39] en comptes d´utilitzar el que proporciona Modeller. El numero de models obtinguts en aquest és 89563 (cal notar que hi ha un cert guany de models, ja que Mammoth és capaç d´alinear certes parelles *target-template* que Modeller per si sol no podia). De la mateixa manera que en cas anterior només hem considerat aquells models pels quals tots els àtoms de la cavitat més gran estan alineats, amb el que el número final de models emprats fou 53507.

Finalment varem emprar un control sobre la metodologia de modelat en condicions ideals, per tal d´avaluar el límit superior de qualitat esperat. El control fou constituït per 3802 auto-models corresponents als 3802 dominis utilitzats en l´estudi.

IDENTITAT DE SEQÜÈNCIA

Les identitats de seqüència han estat calculades tal i com ho fa el programa Modeller: #residus iguals/#residus de la seqüència més curta.

És important mencionar que donat un *target* i un *template*, la identitat calculada és la global (tota la seqüència), i no la corresponent a la part alineada en l´etapa de modelat.

CÀLCUL DE CAVITATS

El càlcul de cavitats s´ha efectuat amb el programa Surfnet [40]. Aquesta aplicació dóna, per una determinada proteïna, un seguit de cavitats, definides per un conjunt d´àtoms. Com es detallarà més endavant, al llarg d´aquest treball ens hem centrat en la cavitat més gran, sovint relacionada amb la funció de la proteïna [40, 41].

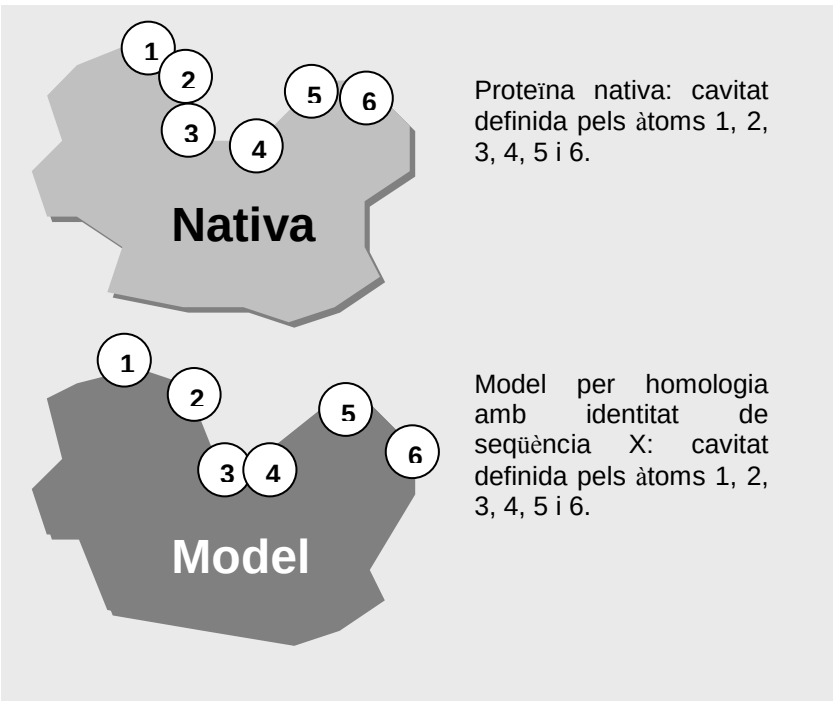
CANVIS EN LES CAVITATS

Hem utilitzat 6 paràmetres per caracteritzar els canvis en les cavitats:

- RMSD (root-mean-square deviation)
- RMSD100 (root-mean-square deviation normalitzat)
- GDT (Global Distance Test)
- Cx (Protrusion Index)
- ΔASA (variació a l`àrea de superfície accessible)
- ΔCN (variació en el número de contactes)

El càlcul de cavitats s´ha efectuat només sobre les proteïnes natives, i no sobre cadascun dels models, és a dir, les cavitats dels models no estan calculades sobre ells, sinó que són les calculades sobre la proteïna nativa. Per tant, si per una determinada proteïna una cavitat està definida pels àtoms a_1, a_2, a_3,... a_n, quan calculem per exemple l´RMSD d´aquesta cavitat entre el model generat i la proteïna experimental ho farem en base aquest conjunt d´àtoms. En la figura 7.2 es pot veure de manera gràfica.

Figura 7.2. Definició de les cavitats a les proteïnes natives i als models per homologia. El càlcul de cavitats s´efectua sobre les natives; els àtoms que determinen la cavitat al model seran els que s´han calculat sobre la nativa.



A continuació es comenten els sis paràmetres estudiats:

a) RMSD: proporciona una idea dels canvis geomètrics experimentats per les estructures comparades. S´ha calculat utilitzant la llista d`àtoms que defineix cada cavitat; d´aquesta manera obtenim un valor numèric que ens diu com de diferent és la cavitat en el model respecte la cavitat original. En alguns casos (figura 7.6) hem

obtingut l´RMSD utilitzant tots els carbonis alfa de la proteïna, per tal de relacionar la qualitat de la cavitat amb la qualitat global del model com es comentarà més endavant.

b) $RMSD_{100}$: l´RMSD és una variable els valors de la qual depenen de la quantitat de punts comparats; així per exemple obtenir un RMSD de 4Å quan comparem dues estructures de 10 residus equival a dir que aquestes són molt diferents; en canvi si obtenim un RMSD de 4Å quan comparem dues estructures de 100 residus podem dir que són similars. Aquesta dependència es pot corregir emprant una normalització de l´RMSD respecte una mida concreta, per exemple 100 residus. Aquesta normalització es fa emprant la següent expressió: RMSD / [1 + 0.5ln(N/100)], on RMSD és l´RMSD estàndard, i N és el número de residus o punts utilitzats en el seu càlcul [42].

c) GDT: és una mesura utilitzada per comparar dues estructures, i permet identificar subestructures comuns entre elles [43]. Correspon al percentatge d`àtoms alineats que es troben per sota un llindar, habitualment 1Å, 2Å, 4Å i 8Å. L´algorisme que es segueix en el nostre estudi es pot consultar al capítol de *Metodologia General*.

d) Cx: és un paràmetre que proporciona una visió local de l´entorn atòmic d´un àtom. Correspon al quocient entre el volum lliure i l´ocupat dins una esfera de 10Å centrada en cadascun dels àtoms pesats. Els seus valors varien entre 0 i 15, corresponent els valors propers a 15 àtoms que sobresurten, i que per tant podrien estar implicats en interaccions proteïna-proteïna. Els càlculs de Cx han estat realitzats amb el programa realitzat per Pintar i col.laboradors [44]. Per aquesta variable es va obtenir un control específic, ja que per un determinat àtom, quan comparem valors de Cx entre estructures podem trobar petites fluctuacions que probablement no tenen significat. Per tal d´establir un llindar per sobre del qual les variacions de Cx són rellevants hem comparat un conjunt de parells de rèpliques de la mateixa estructura, obtinguda sota condicions experimentals diferents. Aquesta llista de parells es va obtenir agrupant totes les estructures PDB [45] de la versió de 25 de Maig de 2007. Tot seguit varem aplicar un seguit de filtres per excloure: models teòrics, residus modificats, residus

incomplets, desconeguts, condicions experimentals extremes i mutants. Les proteïnes que van passar aquests filtres van ser agrupades de nou utilitzant Cd-hit [46], i es van eliminar aquells casos on les mides eren diferents. El total de parells de proteïnes equivalents fou 223. Per cada àtom es va calcular la diferència de Cx entre les rèpliques, i es va trobar que més del 99% de valors de Δcx es trobaven entre -1 i 1. En base això, i aplicat al nostre estudi, per una cavitat concreta s´ha calculat el percentatge d`àtoms pels quals Δcx és més gran a |1|.

e) ΔASA: el càlcul d´accessibilitat atòmica va ser realitzat amb el programa NACCESS [47], amb una sonda de radi 1.4Å. L´accessibilitat és un descriptor que dóna una idea de la capacitat de l´àtom i residus per interactuar amb el medi; així per exemple àtoms amb accessibilitat reduïda tindran tendència a trobar-se el nucli de la proteïna, per tant la interacció amb l´entorn serà mínima.

f) ΔCN: o canvi en el número de contactes. Dóna una idea de com varia la capacitat d´interacció d´un àtom amb el medi. En el nostre cas ens donarà una idea de com varia la capacitat d´interacció dels àtoms de la cavitat en la proteïna experimental i el model. És un paràmetre derivat de l´anterior. Es pot obtenir utilitzant la següent aproximació [48]: $\Delta CN \sim 0.31 \Delta ASA$.

7.3. RESULTATS

Com s´ha comentat en la metodologia s´ha intentat cobrir tot l´espectre d´identitats entre el *target* i *template*, tot i que ens hem centrat majorment en el rang inferior a 60%.

Els dominis CATH utilitzats cobreixen totes les classes (alfa: 24%, beta: 29%, alfa-beta: 45% i altres: 2%), inclouen membres de 33 arquitectures i 390 topologies; així doncs donen una bona cobertura de l´espai estructural dels dominis de proteïnes.

Seguidament es discuteixen els canvis en les cavitats en base a la variació dels paràmetres comentats a la *Metodologia*.

CANVI EN LA FORMA DE LES CAVITATS

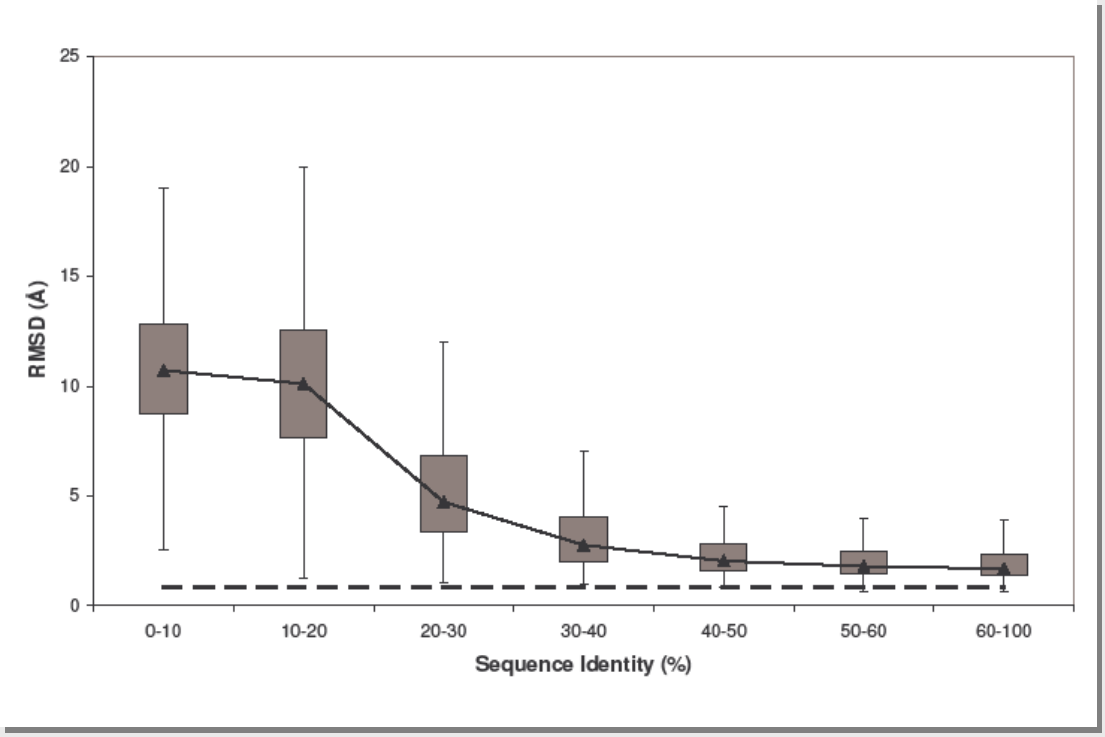
Tal i com s´ha comentat a l´apartat de *Metodologia* s´han emprat sis variables: RMSD, $RMSD_{100}$, GDT, cx, ΔASA i ΔCN.

RMSD

L´RMSD entre l´estructura experimental i el model van ser calculats utilitzant les cavitats definides sobre la primera; més concretament considerant només la cavitat més gran, i que presentava tots els àtoms en la zona alineada en la construcció del model. Com a control es va construir un seguit d´automodels (model per homologia del domini utilitzant el mateix domini com a *template*); els resultats d´aquest control proporcionen una línea basal corresponent als límits del propi programa de modelatge –Modeller–, és a dir, ens donen una idea de l´error introduït pel propi programa en el pas de construcció del model.

A la figura 7.3 es pot veure la variació del RMSD respecte la identitat de seqüència entre *target* i *template*.

Figura 7.3. Relació entre RMSD de les cavitats i SEQ ID. En línia discontinua el control.



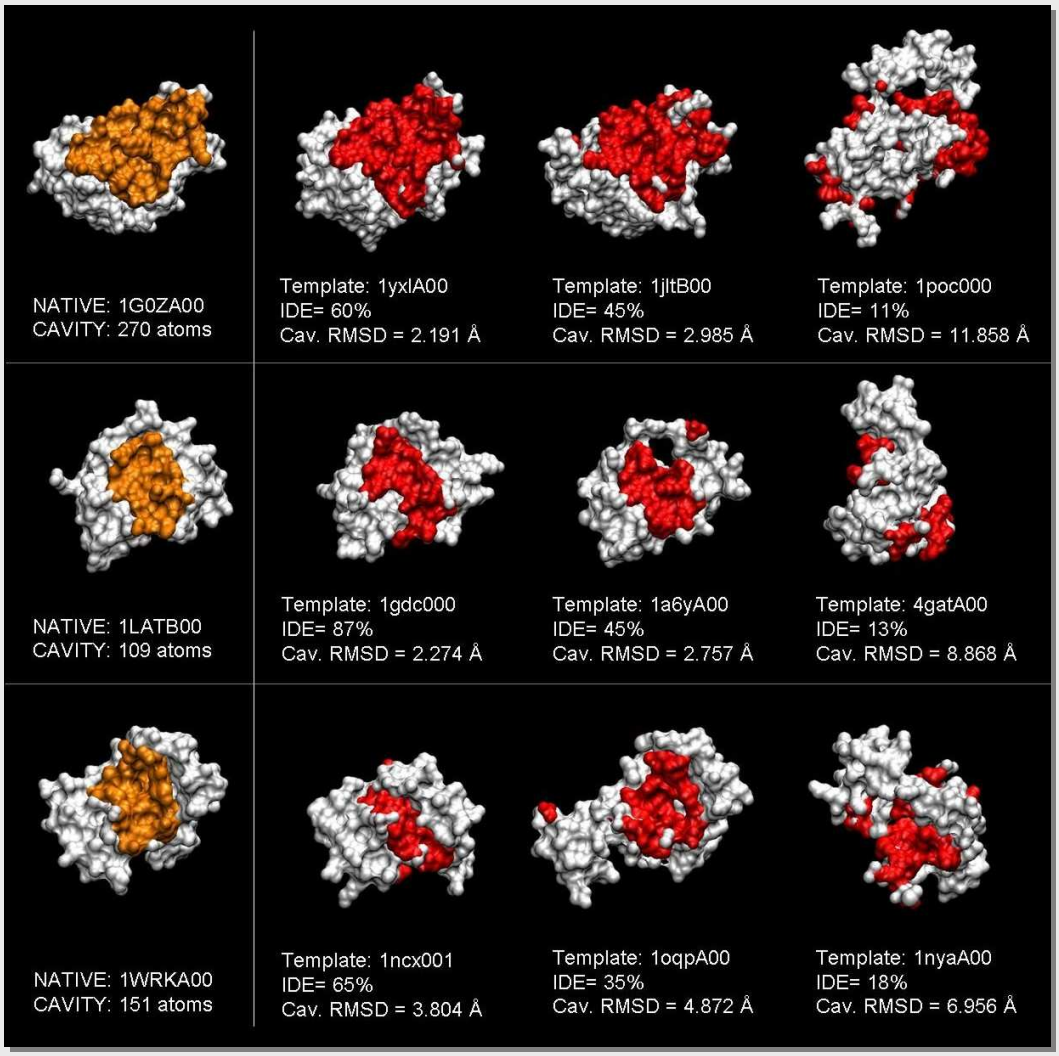
Tal i com es pot esperar, observem que els valors d´RMSD disminueixen a mesura que puja la identitat; tendint de forma asimptòtica cap als valors d´automodelat. Així per exemple, la majoria de models que es troben per sota del 20% d´identitat de seqüència presenten cavitats força distorsionades (més del 75% dels casos amb RMSD superior a 7-8Å). A mesura que la identitat puja ho fa la qualitat de les cavitats, i per exemple per sobre de 30% la gran majoria presenta RMSDs inferiors a 5Å. Per sobre de 40% la tendència s´estabilitza, i s´assoleixen RMSDs propers a 2Å. Apart de la tendència comentada podem observar que:

- A rangs d´identitats propers al 30% hi ha casos on les cavitats modelades presenten RMSDs baixos. Aquesta observació ens indica que possiblement

aquests casos podrien ser emprats en *screening* d´unió de possibles fàrmacs, estudi de funció,...

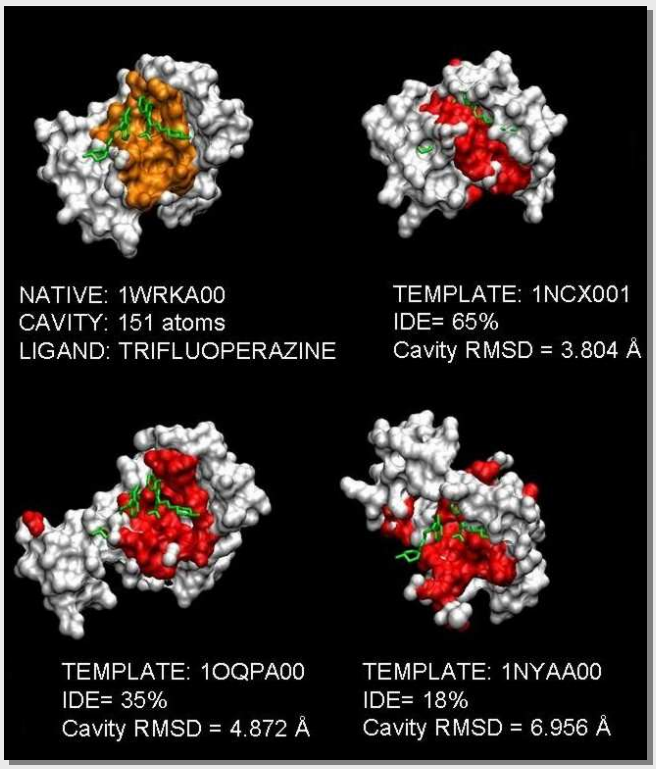
- A rangs d´identitats superiors al 40% és difícil assolir RMSDs propers al control. En aquests casos podria ser convenient utilitzar mètodes d´optimització com dinàmica molecular [49], modelat específic del centre actiu emprant altres *templates* [21], o fins i tot determinació experimental de l´estructura.

Figura 7.4. Exemples de la relació entre RMSD i Seq.id. 1GOZ: Enterotoxina *S. aureus*. 1LAT: Domini d´unió a DNA del receptor de glucocorticoid de *R. norvegicus*. En taronja i vermell es mostren les cavitats en les estructures natives i models per homologia respectivament.



A la figura 7.4 es poden veure dos exemples de com varia la cavitat a la proteïna experimental (en taronja), en relació a tres models obtinguts a diferents identitats de seqüència (en vermell); s´observa visualment i numèrica com a mesura que la identitat de la proteïna experimental respecte el *template* decreix, així ho fa la qualitat de la cavitat.

Figura 7.5. Unió lligand-cavitat en models de diferents qualitats. Es pot veure com la cavitat de la proteïna experimental (taronja) queda alterada en tots tres models (vermell), i amb això els punts d´interacció amb el lligand.



Com s´ha comentat, un dels camps que més benefici pot obtenir del modelat per homologia és el disseny de fàrmacs; ara bé, cal tenir present que si el model construït és dolent el centre actiu pot presentar deformacions importants. A la figura 7.5 veiem com varia el centre actiu del domini N terminal de la troponina cardíaca a mesura que la qualitat del model baixa; i com es perden les interaccions i contactes amb un lligand (trifluoroperazina). De fet en aquest cas concret ni tan sols a identitats del 65% es manté la geometria del centre actiu necessària per interaccionar amb el lligand.

Per tal de completar l´estudi anterior es va explorar la relació entre la qualitat de la cavitat i la qualitat global del model. La importància d´aquest punt és gran si tenim en consideració la possibilitat de refinar el model *a posteriori* emprant tècniques com dinàmica molecular. Aquestes tècniques de refinament tenen la particularitat de considerar tots els àtoms de la mateixa forma, el que pot conduir al següent problema: imaginem un cas on la cavitat d´un model té RMSD baix, i el model globalment té un

RMSD alt; si intentem refinar-lo emprant tècniques de dinàmica molecular segurament provocarem una degradació de la cavitat en detriment d´una millora global [26, 50].

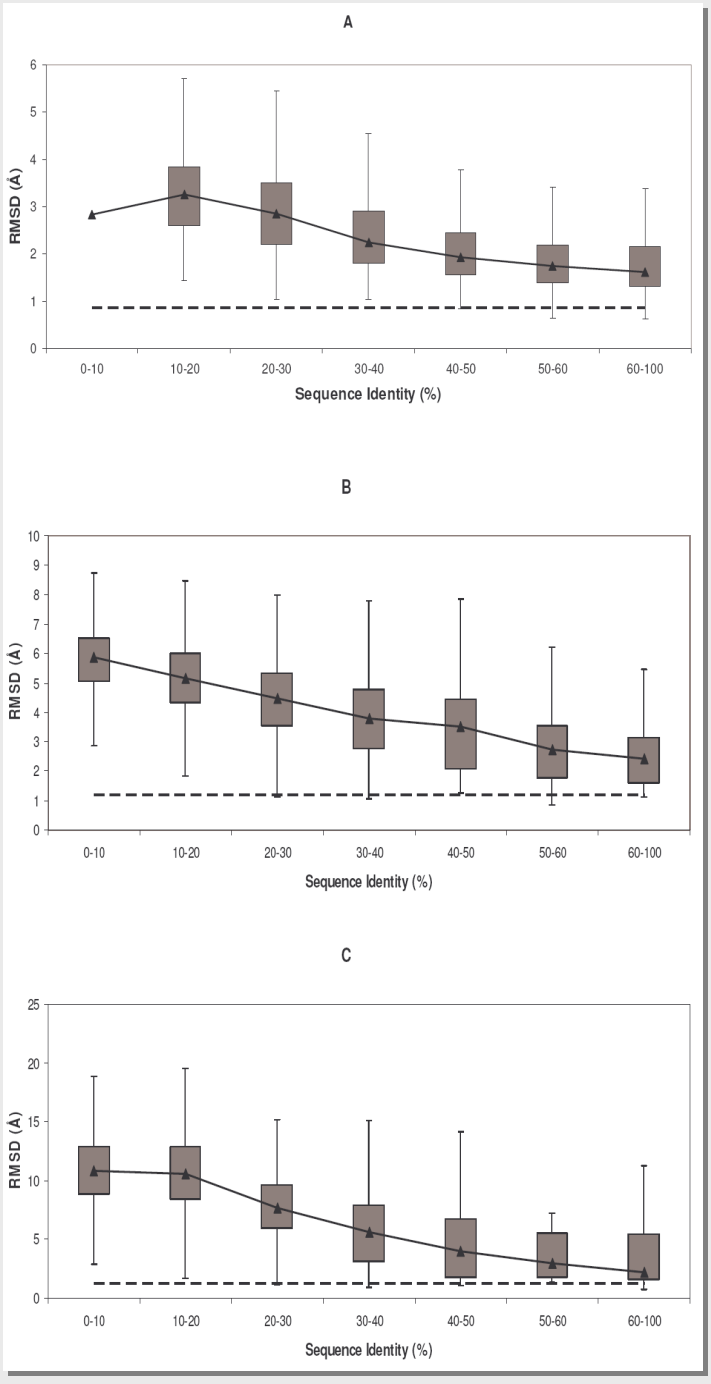


Figura 7.6. Qualitat cavitats en funció de la qualitat de l´esquelet del model. La figura A correspon a aquells models on l´RMSD de l´esquelet del model és de 0 a 3Å; la figura B correspon a aquells models on l´RMSD de l´esquelet del model és de 3 a 6Å; i la figura C aquells models on l´RMSD és superior a 6Å.

Tenint en compte aquesta consideració, varem dividir les dades d´RMSD anteriors (figura 7.3) en tres classes, en base a la qualitat de l´esquelet del model (RMSD de carbonis alfa): models alta qualitat (0Å – 3Å), models de mitjana qualitat (3Å – 6Å) i models de baixa qualitat (més de 6Å). En la figura 7.6 es poden apreciar els resultats. Globalment podem dir que per sobre el 40% d´identitat una proporció considerable de cavitats presenten un RMSD inferior al que presenta l´esquelet del model; sobretot si ens centrem en els casos on aquest esquelet presenta alta i mitjana qualitat (figures 7.6A i 7.6B). Aquest comportament es pot entendre si considerem dos factors aparentment oposats: i) existència de restriccions funcionals fan que les cavitats estiguin més ben modelades; ii) la presència de parts mal modelades al llarg de la proteïna té l´efecte oposat en l´esquelet de la proteïna.

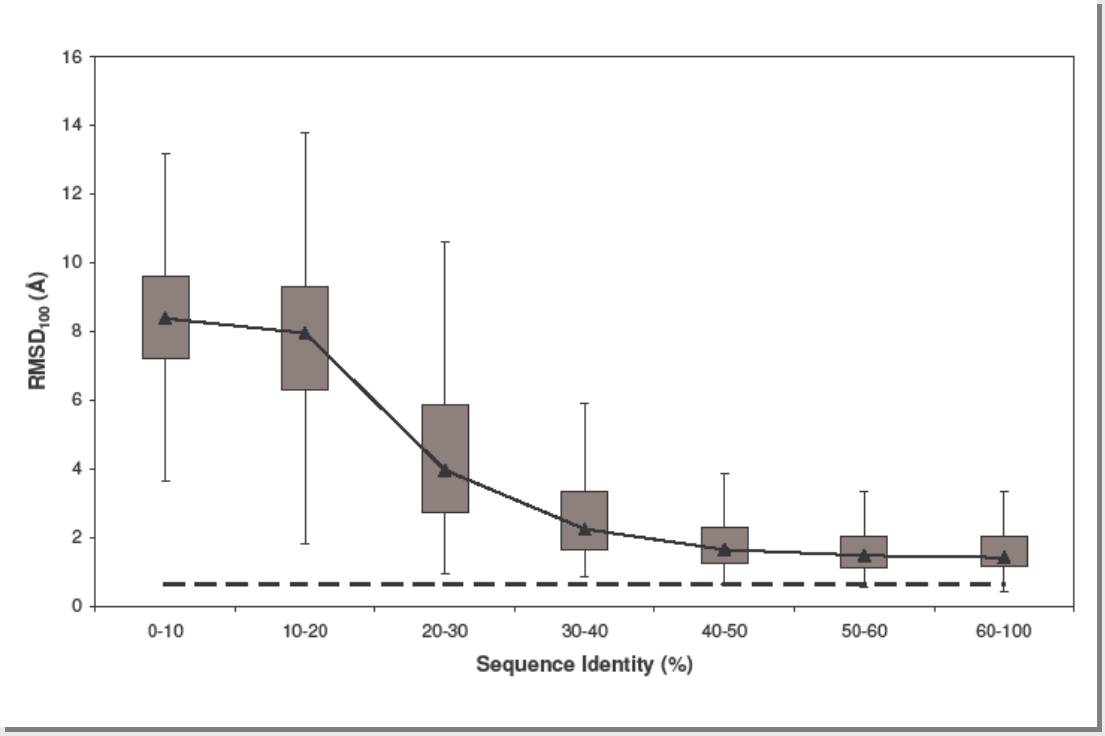
Així doncs, tornant a la possibilitat d´un refinament *a posteriori*, i focalitzant-nos en el rang de dades amb identitats superiors al 40%, aquest requeriria possiblement algun tipus de restricció dels àtoms que conformen la cavitat, per tal d´evitar que amb la millora global del model la cavitat empitjori. Pel que fa als models amb identitats inferiors al 30% ens trobem amb un escenari completament oposat, i sí que podríem pensar en utilitzar refinaments per dinàmica molecular sense massa risc de comprometre la qualitat de la cavitat.

$RMSD_{100}$

Tenint en compte que les cavitats estudiades en aquest treball tenen diferents mides es va decidir emprar l´$RMSD_{100}$, que com ja s´ha comentat és un RMSD normalitzat, i permet expressar tots els canvis d´RMSD observats en una mateixa escala.

El comportament observat per a l´$RMSD_{100}$ (figura 7.7) és comparable a l´obtingut per l´RMSD (figura 7.3): mateixa tendència asimptòtica cap als valors d´automodelat. Això confirma la independència dels resultats principals respecte la mida de la cavitat.

Figura 7.7. Relació entre RMSD100 i la identitat de seqüència. La línia discontínua representa el control.



GDT

Com s´ha comentat a la *Metodologia* el GDT és una mesura directament relacionada amb la presència de subestructures de bona qualitat, i treballa identificant el percentatge d´àtoms modelats per sota un llindar donat [43]. Per al seu càlcul hem utilitzat els llindars habituals (1Å, 2Å, 4Å i 8Å).

En els resultats obtinguts (figura 7.9) veiem dos escenaris depenent de si la identitat de seqüència és major o menor al 30%. Per sobre del 30% una proporció important de cavitats mostren valors de GDT elevats, fins i tot en llindars més restrictius (GDT_1 i GDT_2), indicant l´existència de subestructures de qualitat alta. Si ens fixem en les zones d´identitat de seqüència inferiors al 30% podem veure com en les figures 7.9A i 7.9B (GDT_1 i GDT_2) els valors de GDT són baixos, per tant el número de subestructures de bona qualitat és força reduït, contràriament al que passa en la figura

7.9C (GDT_4), on podem veure que hi ha una fracció de subestructures de qualitat relativament important.

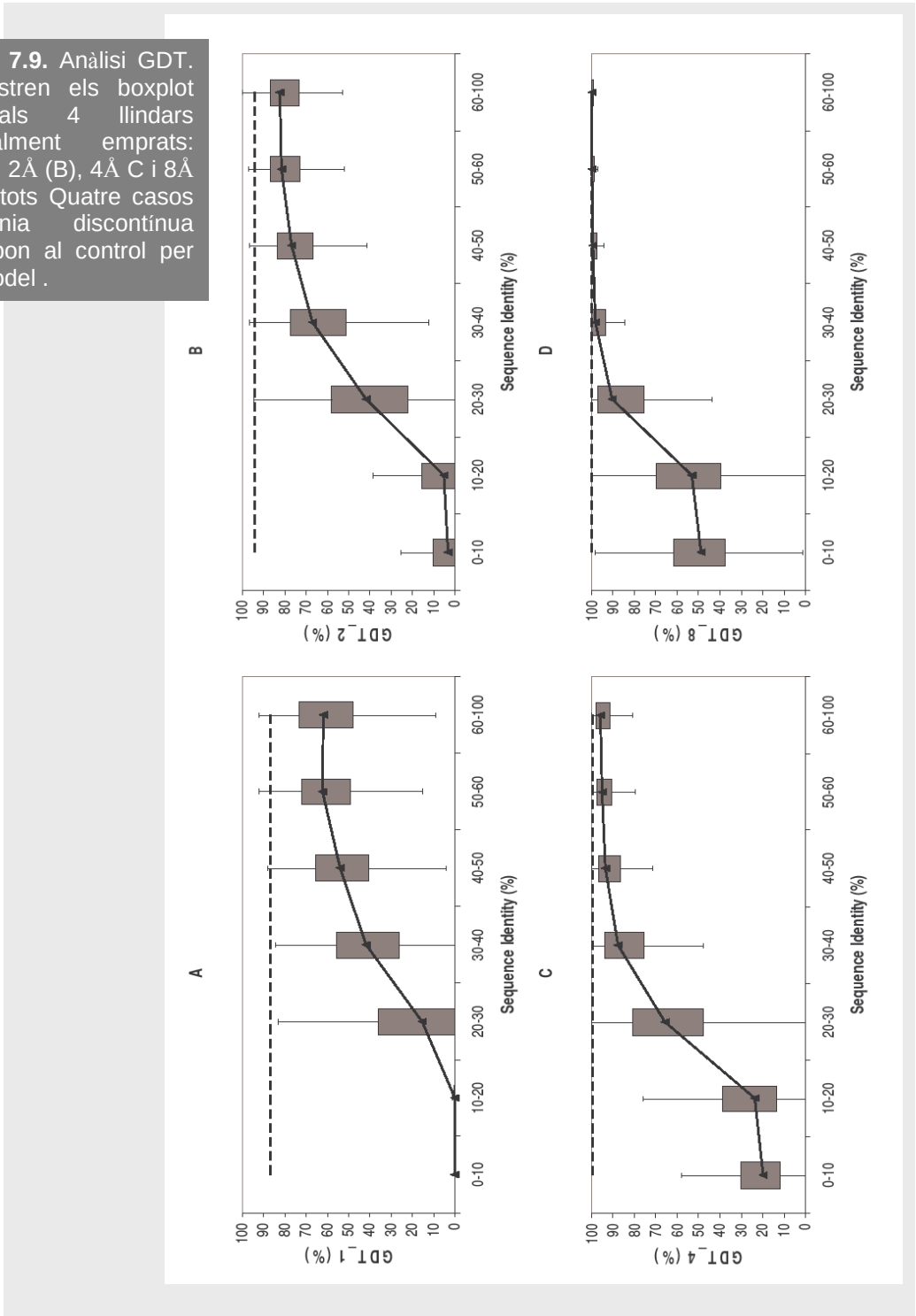


Figura 7.9. Anàlisi GDT. Es postren els boxplot per als 4 llindars habitualment emprats: 1Å (A), 2Å (B), 4Å C i 8Å D. En tots Quatre casos la línia discontínua correspon al control per automodel .

Tot i que aquestes subestructures no poden emprar-se per aplicacions com disseny de fàrmacs, o altres aplicacions que requereixin una qualitat estructural elevada, sí que poden ser emprades com a punt de sortida per un refinament posterior del model o de la cavitat [17].

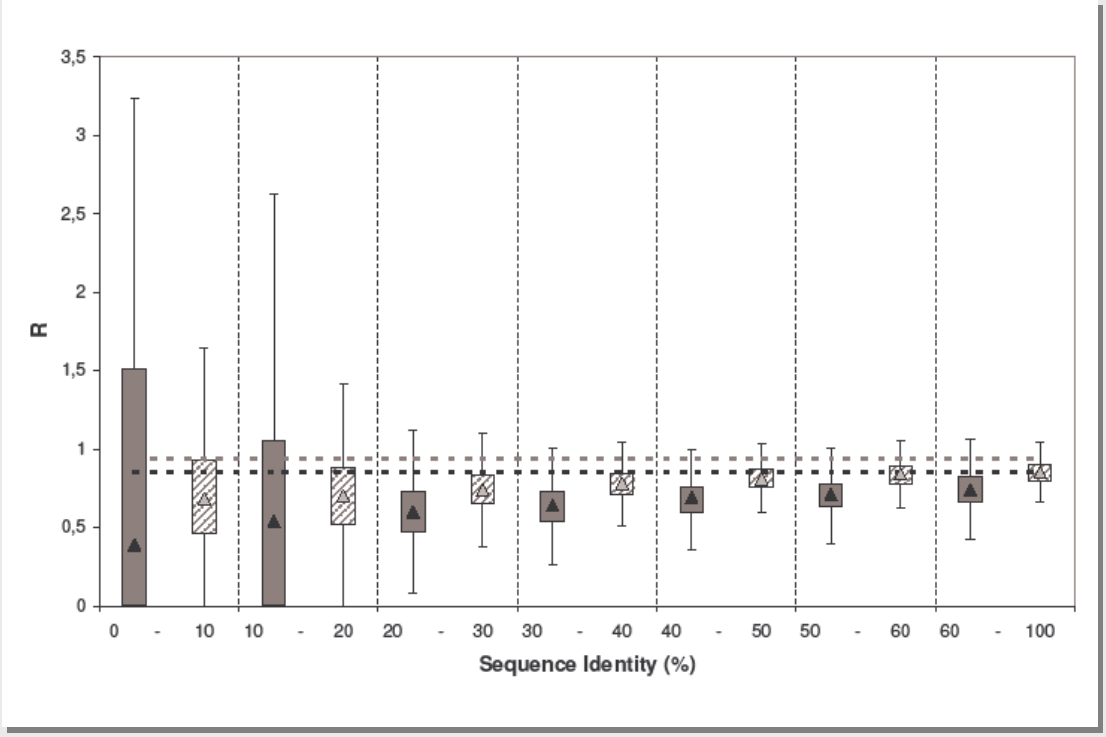
Es decidí refinar l´anàlisi estudiant més detalladament el comportament dels àtoms de les cadenes laterals; la raó és que aquests àtoms constitueixen una fracció important dels àtoms d´una cavitat, i normalment és complicat modelar-los [18]. Concretament per cada cavitat hem determinat un coeficient R:

R = %àtoms cadena lateral a un GDT concret / % àtoms cadena lateral de la cavitat

Si els àtoms de les cadenes laterals estan modelats de forma similar als àtoms de l´esquelet R tendirà a 1. No obstant si les cadenes laterals estan modelades pitjor que la cadena principal R serà menor a 1, i viceversa. Els resultats per a GDT_1 i GDT_2 (aquests llindars són els que identifiquen subestructures de qualitat elevada) es mostren a la figura 7.8.

Si ens fixem en les distribucions de R observem primer de tot que els valors de l´automodel són lleugerament inferiors a 1; això indica que fins i tot en una situació ideal de modelatge les cadenes laterals dels residus són modelades de forma més pobre que la resta de l´esquelet. Una segona observació és que en general els valors de R són inferiors a 1, tot i que tendeixen asimptòticament cap als valors del control a mesura que puja la identitat de seqüència. Per tant podem concloure que els àtoms de la cadena principal tenen una major contribució a les parts més ben modelades, no obstant això, a mesura que puja la identitat de seqüència, també ho fa la contribució de les cadenes laterals.

Figura 7.8. Contribució de les cadenes laterals a la qualitat de les cavitats. Es postren les distribucions corresponents a GDT_1 (gris) i GDT_2 (ratllat). Les línies ratllades corresponen als valors control de l´automodel: gris fosc per GDT_1 i gris clar per GDT_2.



Les grans fluctuacions observades entre els rangs d´identitat de 0 a 30%, en particular a GDT_1, molt possiblement siguin conseqüència d´un modelat poc acurat, possiblement aleatori, de les cadenes laterals.

CX

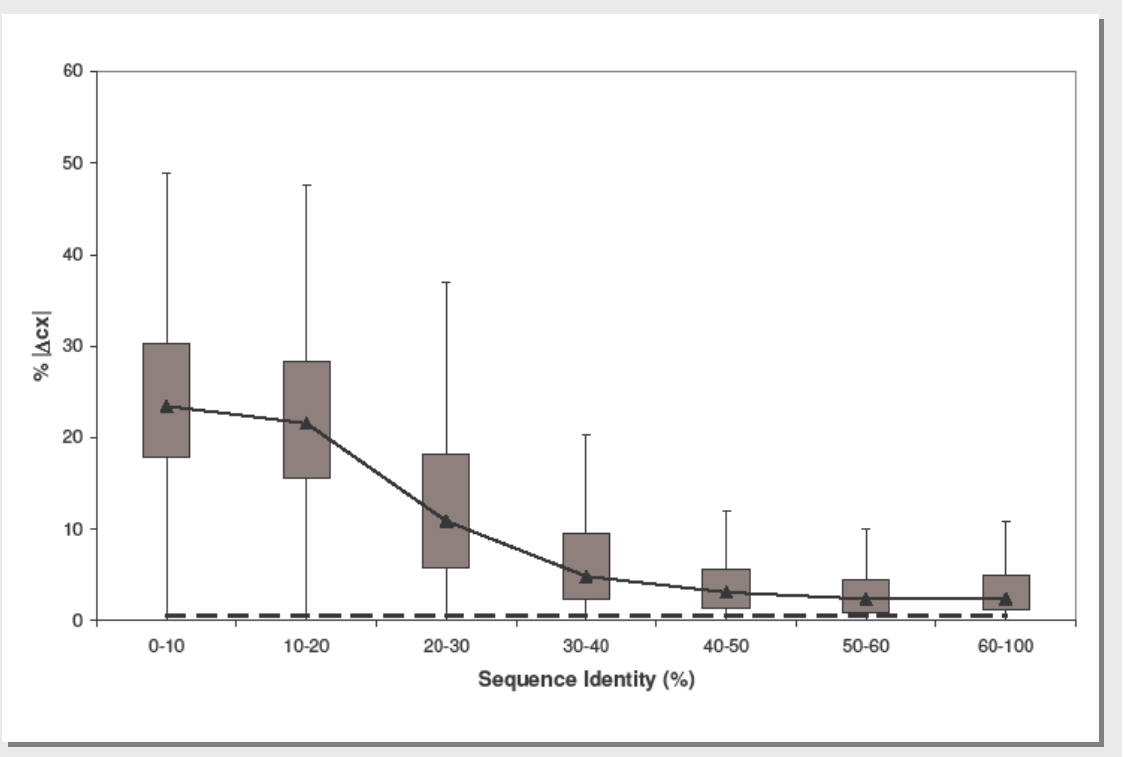
Com s´ha comentat a l´apartat corresponent de *Metodologia*, CX és un quocient de volums que dóna una mesura local de l´entorn d´un àtom [44]. En aquest treball, per cada cavitat estudiada hem calculat el percentatge d`àtoms on el valor de cx variava entre el model construït i la proteïna experimental, i s´ha examinat la dependència d´aquesta variació respecte la identitat de seqüència.

Tal i com està explicat als mètodes, per establir un llindar per sobre del qual considerem que cx varia, hem agafat 223 de parells d´estructures experimentals, corresponents a la mateixa proteïna, però determinades en diferents condicions. Fet això hem calculat les diferències entre àtoms equivalents i s´ha analitzat la distribució

de variacions. El resultat és que en un 99% dels casos la diferència en el valor cx es troba entre 1 i -1; així doncs podem considerar que qualsevol diferència que surti d´aquest interval no és fruit del soroll. En base això, el percentatge d`àtoms d´una cavitat pels quals cx varia correspon a aquells àtoms on la diferència de cx entre proteïna experimental i model és superior a |1|.

Pel que fa als resultats (Figura 7.10) són consistents amb els mostrats per valors d´RMSD i GDT: per sobre de 30-40% d´identitat el percentatge d´àtoms amb variacions de cx és reduït, tendint als valors de l´automodel. Per sota de 30-40% aquest percentatge puja notablement, mostrant una transició similar a la observada en els estudis d´RMSD (figura 7.3). Per tant podem concloure que a identitats baixes les cavitats experimenten canvis tant a nivell global (RMSD) com a nivell local (cx).

Figura 7.10. Convervació de cx. Es mostra la distribució en relació a la seq.id. del percentatge d`atoms de la cavitat amb diferències de cx superiors a |1|. La línia discontínua correspon als valors control de l´automodel



ΔASA I ΔCN

En aquest darrer apartat es discuteixen els canvis en la superfície accessible atòmica (ASA) i número de contactes (CN) dels àtoms de la cavitat major.

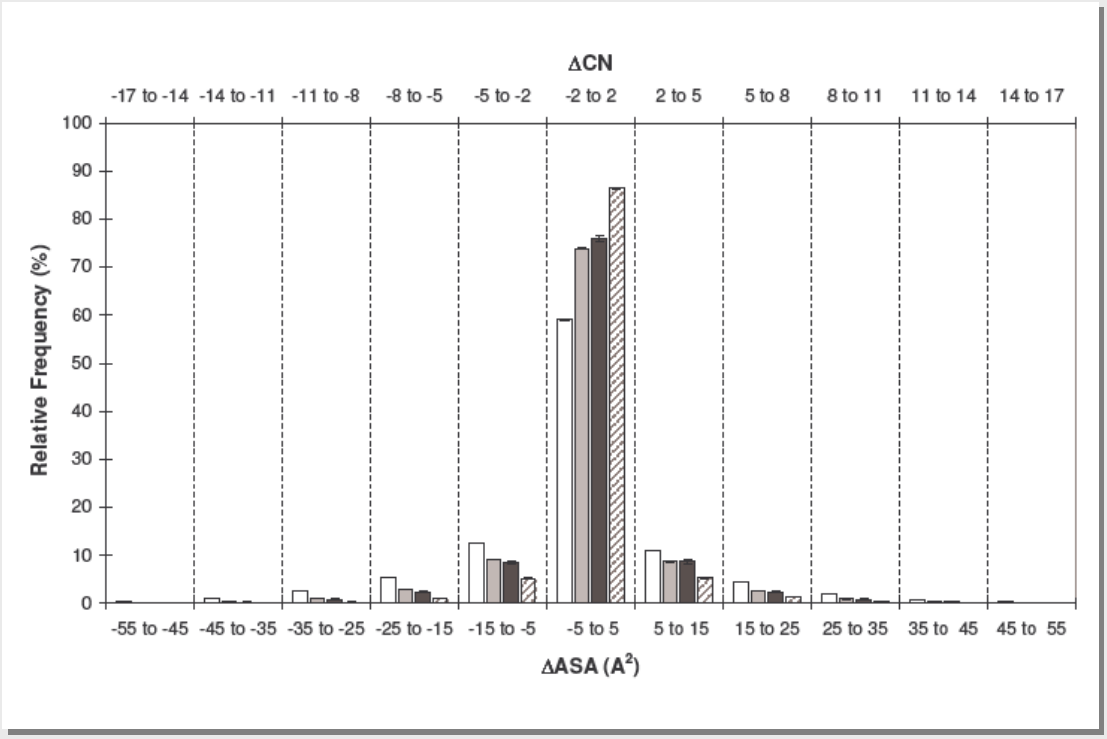
Per tal d´estudiar com varia l´ASA i CN amb la identitat de seqüència s´ha dividit el conjunt de dades en base aquesta última, obtenint tres grups: baixa identitat (<30%), mitjana identitat (30-60%) i alta identitat (>60%). Per cadascun d´aquests grups s´ha calculat la variació d´ASA dels àtoms que formen la cavitat més gran ($\Delta ASA = ASA_{experimental} - ASA_{model}$). A la figura 7.11 es poden comprovar els següents punts, que lliguen amb les dades obtingudes fins ara:

- Les variacions d´ASA tendeixen cap a valors d´automodel a mesura que la seq.id augmenta.
- Les variacions d´ASA creixen a mesura que la qualitat del model baixa.
- Hi ha una diferència important d´ΔASA entre models de baixa qualitat i models de mitjana/alta qualitat.
- Les variacions d´ASA s´agrupen al voltant de 0; això indica que el protocol de modelatge té poc efecte sobre la variació d´accessibilitat dels àtoms.

Certes aplicacions del models per homologia, com podrien ser disseny de fàrmacs, o estudis d´interaccions enzim substrat, requereixen un modelat acurat de les interaccions atòmiques entre el model i el possible substrat. Per tal de proporcionar una estimació de com poden variar aquestes interaccions segons la qualitat del model hem utilitzat un paràmetre addicional derivat d´ASA: canvi en el número de contactes o ΔCN. Aquest paràmetre s´obté de la següent aproximació proposada per Colonna-Cesari i Sander [48]: $\Delta CN \sim 0.31\Delta ASA$. La variació de CN dóna una idea aproximada de com els canvis en l´accessibilitat d´una cavitat modifiquen la capacitat d´aquesta per establir interaccions amb altres molècules.

Els resultats (Figura 7.11) ens mostren que fins i tot per models de baixa qualitat els àtoms de les cavitats presenten variacions de CN al voltant de 3; això vol dir que aquests àtoms han guanyat o perdut l´habilitat d´establir tres interaccions en promig. Aquesta mateixa situació és comparable pels models de qualitat mitjana.

Figura 7.11. Conservació ASA i CN. En la figura es mostra a l´eix X inferior la distribució de ΔASA (ΔASA = ASAexperimental – ASAmodel) pels àtoms de les cavitats en 4 casos: identitat <30% (blanc); identitat 30-60% (gris clar); identitat >60% (gris fosc) i control (ratllat). A l´eix X superior es mostra ΔCN.



FACTORS QUE AFECTEN A LA QUALITAT DE LA CAVITAT

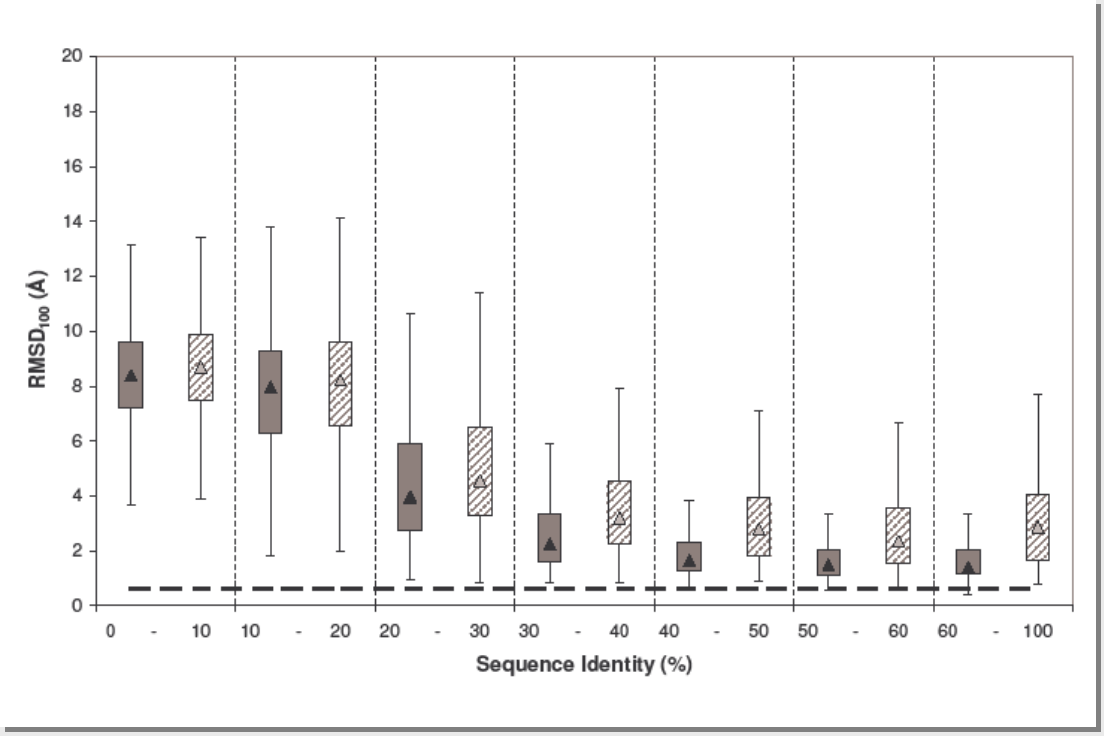
Per concloure el treball s´ha estudiat l´efecte de diversos factors sobre la qualitat de les cavitats; més concretament ens hem centrat en quatre punts:

- Efecte dels residus no alineats en les cavitats.
- Millora sobre les cavitats quan s´utilitza un alineament òptim.
- Influència de famílies estructurals superpoblades.
- Extrapolació dels resultats a les cinc primeres cavitats

Efecte dels residus no alineats

Alhora de construir un model per homologia, les zones no alineades condueixen a regions pitjor modelades que la resta [19, 20, 51]. És per això que en els estudis realitzats aquí ens hem centrat en aquelles cavitats els àtoms de les quals es troben alineats. No obstant això, es poden donar casos on el número de residus no alineats sigui prou petit com per que les restriccions imposades pels residus alineats siguin suficients per general un model acceptable.

Figura 7.12. Efecte de residus no-alineats. En gris es postren els valors de RMSD100 per a les cavitats on tots els àtoms estan alineats. En ratllat es postren les cavitats per les quals al menys un 75% d`àtoms estan alineats



Per tal d´explorar aquesta idea hem ampliat l´estudi incloent totes aquelles cavitats que presentaven una petita fracció d`àtoms no alineats (<= 25%). La figura 7.12 mostra la comparació d´aquestes cavitats amb aquelles on el 100% d`àtoms estaven alineats (cal remarcar que són dades d´$RMSD_{100}$, per tenir en compte un possible efecte de mida). La tendència és similar en ambdós casos, tot i que els valors d´$RMSD_{100}$ de les cavitats completament alineades són menors per sobre del 30%

d´identitat. Així doncs podem concloure que les cavitats que no estan completament alineades poden tenir certa utilitat, ja que no es degraden de forma massa accentuada amb la variació d´identitat; si més no, no ho fan gaire més que aquelles on tots els àtoms estan alineats.

Millora de les cavitats emprant alineaments òptims

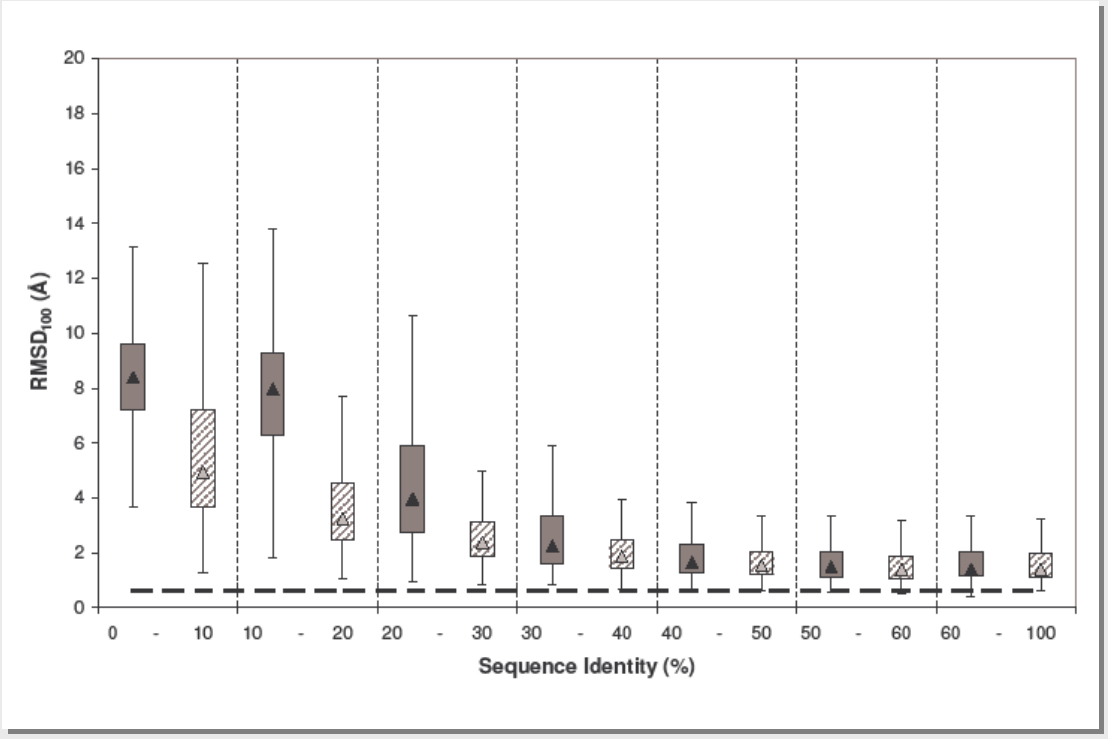
Tenint en compte que en el modelat per homologia el pas crític és l´alineament entre *target* i *template*, hem intentat establir la màxima qualitat que pot ser aconseguida millorant aquest alineament. Aquest punt té particular interès ja que pot ajudar a decidir si val la pena invertir temps i esforç millorant un alineament *target–template*.

Hem emprat alineaments estructurals, que són els millors alineaments que es poden obtenir entre dues seqüències. Per això hem reconstruït els models per homologia no emprant els alineaments que generava el mateix Modeller, sinó emprant alineaments obtinguts amb el programa Mammoth [38, 39]. Sobre aquests models hem comparat la cavitat major de la proteïna nativa (amb el 100% d´àtoms alineats); exactament igual que s´ha fet amb els models per homologia tradicionals. Les dades dels valors d´$RMSD_{100}$ comparant les dades per l`alineament òptim amb les originals es mostren a la figura 7.13.

Podem apreciar dos escenaris segons la identitat de seqüència. En el primer veiem que per identitats de seqüència inferiors al 30% les cavitats derivades dels models generats per alineament estructural són clarament millors a les dels models obtinguts per alineaments de seqüència estàndards. Així doncs en casos on la identitat de seqüència entre *target* i *template* és baixa, la millora de l´alineament pot ser molt beneficiosa de cara a obtenir models millors, tal i com esperaríem.

L´altre situació és la que s´ofereix a identitats superiors al 30%; en aquest cas no sembla que una millora en l´alineament pugui conduir a un modelatge més acurat de les cavitats.

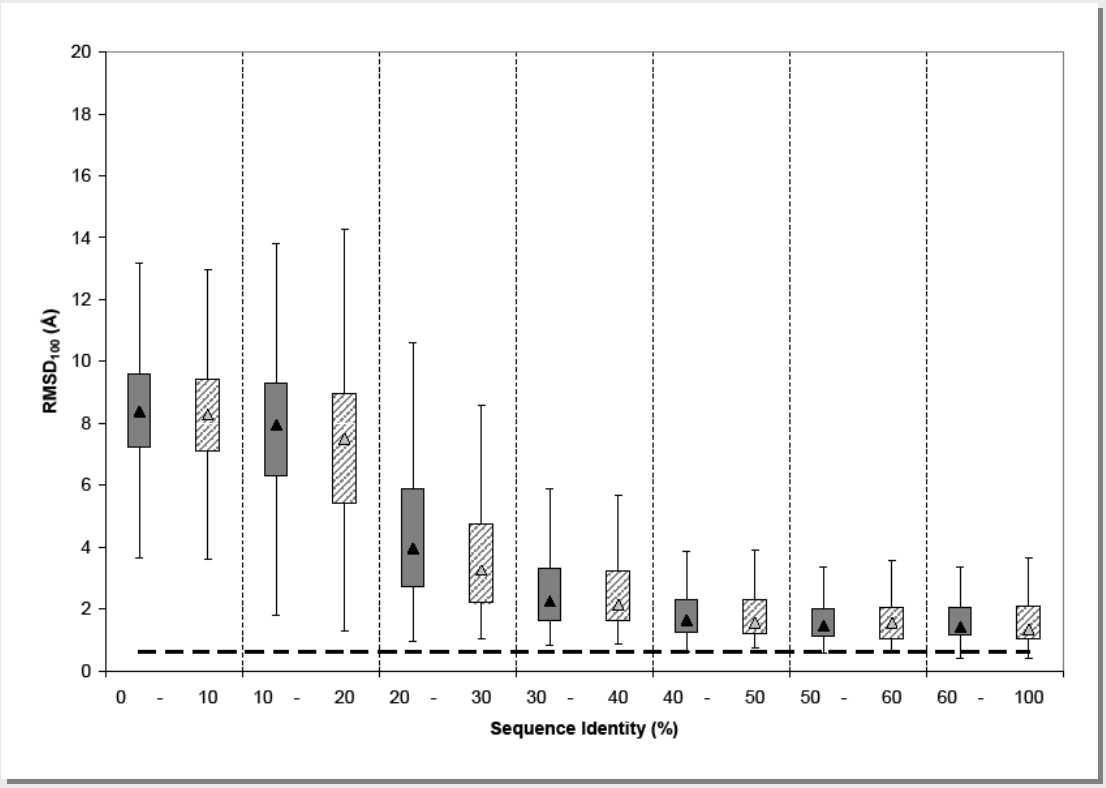
Figura 7.13. Alineament estructural/seqüència. Distribucions dels valors de RMSD100 per a les cavitats obtingudes amb alineament de seqüència (gris fosc) i per a les cavitats obtingudes amb alineament estructural (ratllat). La línia discontínua correspon al control .



Influència de famílies estructurals superpoblades

Hi ha diverses famílies de dominis emprats que contribueixen a un major número de models (ex. Immunoglobulines). Per tal de comprovar que les conclusions generals del treball no estan desviades per aquestes famílies s´han reproduït els resultats mostrats a la Figura 7.7 ($RMSD_{100}$ en funció de seq.id) però només tenint en compte aquelles famílies de proteïnes que contribueixen en menys de 100 models cadascuna. Tal i com es pot veure a la figura 7.14 hi ha poques diferències de tendència entre ambdós conjunts de dades.

Figura 7.14. Distribucions de valors d´RMSD100 corresponent al conjunt de dades original (gris fosc) i al conjunt format per famílies estructurals que contribueixen amb menys de 100 models (ratllat).

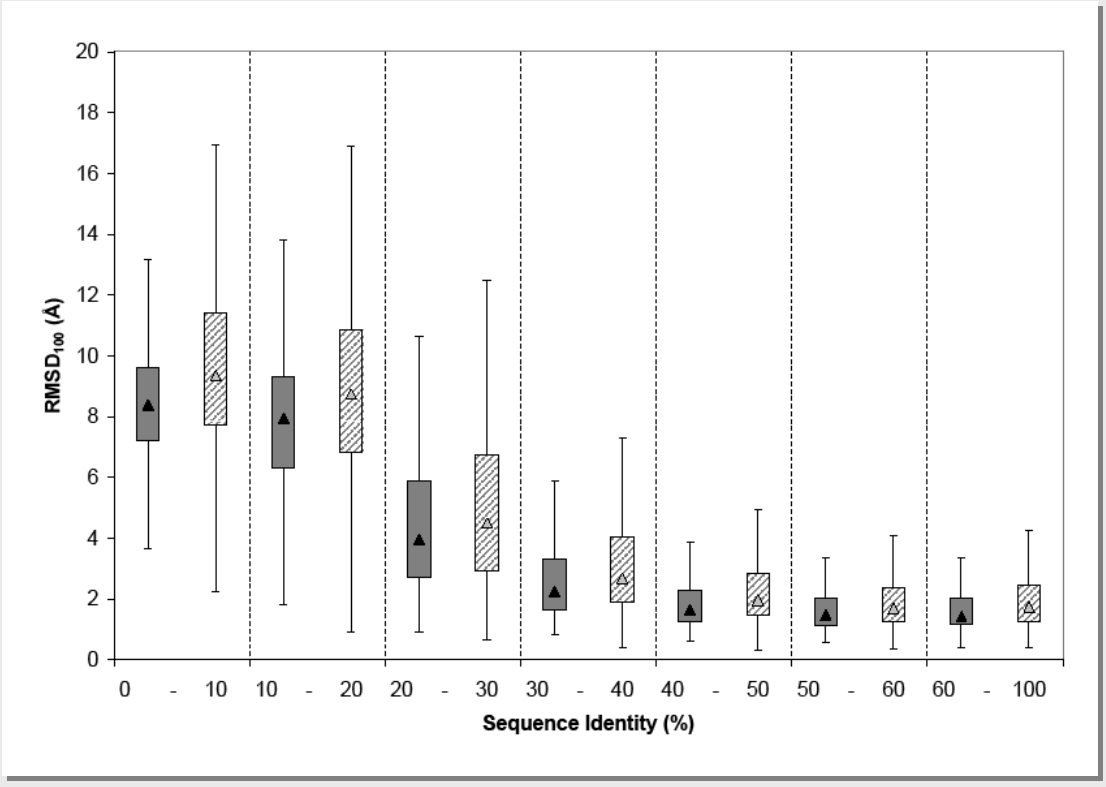


Extrapolació dels resultats a les cinc primeres cavitats

En alguns casos pot succeir que el centre funcional de la proteïna estigui en una cavitat secundària i no a la cavitat més gran; o fins i tot que alguna cavitat secundària correspongui a un centre de regulació al.lostèrica. Per tal d´explorar aquesta possibilitat hem ampliat els resultats obtinguts a les cinc primeres cavitats més grans (amb el 100% dels àtoms alineats). Els resultats es poden mostren a la figura 7.15.

Podem apreciar que les tendències són similars, tot i que els valors d´$RMSD_{100}$ són significativament majors en el cas de les cinc primeres cavitats, sobretot a rangs d´identitat baixes. Això suggereix que les cavitats menors són més difícils de modelar, possiblement per que corresponen a zones menys conservades estructuralment entre homòlegs.

Figura 7.15. Distribució de RMSD100 per a les dades emprant la cavitat més gran (gris fosc) o bé emprant les cinc primeres cavitats més grans (ratllat).



7.4. DISCUSSIÓ

En aquest capítol s´ha presentat un estudi quantitatiu sobre com varia la qualitat de les cavitats en els models per homologia, segons la identitat entre la seqüència *target* i el *template* emprat per construir-los. Concretament les variacions de qualitat s´han analitzat en base a sis variables: RMSD, $RMSD_{100}$, GDT, cx, ΔASA i ΔCN.

Tenint en compte aquests paràmetres, els nostres anàlisis mostren que per sota d´identitats del 20% la qualitat de la cavitat presenta una distorsió important, tant global (figura 7.3) com local (Figures 7.9, 7.10 i 7.11). Aquests resultats suggereixen que tot i que en aquests rangs d´identitat podem ser capaços de crear models útils, la gran majoria tendiran a tenir una qualitat pobre. No obstant això, sempre existeix la possibilitat de sotmetre el model o bé la qualitat a tècniques de refinament, tot i que la millora continuarà passant per l´obtenció d´un bon alineament (Figura 7.13).

Per sobre d´identitats de 30-40% la principal restricció alhora d´obtenir un model bo és la selecció del *template*; com s´ha vist l´estructura global del model tendeix a ser pitjor que l´estructura de la cavitat, possiblement a causa de les restriccions funcionals d´aquesta darrera. En aquest cas, pensant en un refinament, una solució vàlida seria fixar els àtoms de les cavitats, i sotmetre la resta de model a tècniques com dinàmica molecular.

7.5. REFERÈNCIES

1. Vitkup, D., et al., *Completeness in structural genomics.* Nat Struct Biol, 2001. **8**(6): p. 559-66.
2. Marsden, R.L., T.A. Lewis, and C.A. Orengo, *Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint.* BMC Bioinformatics, 2007. **8**: p. 86.
3. Sadreyev, R.I. and N.V. Grishin, *Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds.* BMC Struct Biol, 2006. **6**: p. 6.
4. O'Toole, N., S. Raymond, and M. Cygler, *Coverage of protein sequence space by current structural genomics targets.* J Struct Funct Genomics, 2003. **4**(2-3): p. 47-55.
5. Burley, S.K. and J.B. Bonanno, *Structuring the universe of proteins.* Annu Rev Genomics Hum Genet, 2002. **3**: p. 243-62.
6. Chance, M.R., et al., *Structural genomics: a pipeline for providing structures for the biologist.* Protein Sci, 2002. **11**(4): p. 723-38.
7. Goh, C.S., et al., *Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis.* J Mol Biol, 2004. **336**(1): p. 115-30.
8. Lesley, S.A., et al., *Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline.* Proc Natl Acad Sci U S A, 2002. **99**(18): p. 11664-9.
9. O'Toole, N., et al., *The structural genomics experimental pipeline: insights from global target lists.* Proteins, 2004. **56**(2): p. 201-10.
10. Page, R., et al., *NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1901-5.

11. Peti, W., et al., *Towards miniaturization of a structural genomics pipeline using micro-expression and microcoil NMR.* J Struct Funct Genomics, 2005. **6**(4): p. 259-67.

12. Chandonia, J.M. and S.E. Brenner, *The impact of structural genomics: expectations and outcomes.* Science, 2006. **311**(5759): p. 347-51.

13. Levitt, M., *Growth of novel protein structural data.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3183-8.

14. Spraggon, G., et al., *On the use of DXMS to produce more crystallizable proteins: structures of the T. maritima proteins TM0160 and TM1171.* Protein Sci, 2004. **13**(12): p. 3187-99.

15. Symersky, J., et al., *Structural genomics of Caenorhabditis elegans: structure of the BAG domain.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 9): p. 1606-10.

16. Todd, A.E., et al., *Progress of structural genomics initiatives: an analysis of solved target structures.* J Mol Biol, 2005. **348**(5): p. 1235-60.

17. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.

18. Ginalski, K., *Comparative modeling for protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(2): p. 172-7.

19. Sanchez, R., et al., *Protein structure modeling for structural genomics.* Nat Struct Biol, 2000. **7 Suppl**: p. 986-90.

20. Peitsch, M.C., *About the use of protein models.* Bioinformatics, 2002. **18**(7): p. 934-8.

21. Marti, L., et al., *Exploring the binding mode of semicarbazide-sensitive amine oxidase/VAP-1: identification of novel substrates with insulin-like activity.* J Med Chem, 2004. **47**(20): p. 4865-74.

22. Chakravarty, S. and R. Sanchez, *Systematic analysis of added-value in simple comparative models of protein structure.* Structure, 2004. **12**(8): p. 1461-70.

23. Chakravarty, S., L. Wang, and R. Sanchez, *Accuracy of structure-derived properties in simple comparative models of protein structures.* Nucleic Acids Res, 2005. **33**(1): p. 244-59.

24. Sanchez, R. and A. Sali, *Advances in comparative protein-structure modelling.* Curr Opin Struct Biol, 1997. **7**(2): p. 206-14.

25. Tondel, K., *Prediction of homology model quality with multivariate regression.* J Chem Inf Comput Sci, 2004. **44**(5): p. 1540-51.

26. Tramontano, A., R. Leplae, and V. Morea, *Analysis and assessment of comparative modeling predictions in CASP4.* Proteins, 2001. **Suppl 5**: p. 22-38.

27. Tress, M., et al., *Assessment of predictions submitted for the CASP6 comparative modeling category.* Proteins, 2005. **61 Suppl 7**: p. 27-45.

28. DeWeese-Scott, C. and J. Moult, *Molecular modeling of protein function regions.* Proteins, 2004. **55**(4): p. 942-61.

29. Orengo, C.A. and J.M. Thornton, *Protein families and their evolution-a structural perspective.* Annu Rev Biochem, 2005. **74**: p. 867-900.

30. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of function in protein superfamilies, from a structural perspective.* J Mol Biol, 2001. **307**(4): p. 1113-43.

31. Bray, J.E., et al., *A practical and robust sequence search strategy for structural genomics target selection.* Bioinformatics, 2004. **20**(14): p. 2288-95.

32. Rost, B., *Protein structures sustain evolutionary drift.* Fold Des, 1997. **2**(3): p. S19-24.

33. Lesk, A.M., M. Levitt, and C. Chothia, *Alignment of the amino acid sequences of distantly related proteins using variable gap penalties.* Protein Eng, 1986. **1**(1): p. 77-8.

34. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.

35. Pearl, F.M., et al., *The CATH database: an extended protein family resource for structural and functional genomics.* Nucleic Acids Res, 2003. **31**(1): p. 452-5.

36. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D226-9.

37. Dunbrack, R.L., Jr., *Sequence comparison and protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(3): p. 374-84.

38. Lupyan, D., A. Leo-Macias, and A.R. Ortiz, *A new progressive-iterative algorithm for multiple structure alignment.* Bioinformatics, 2005. **21**(15): p. 3255-63.

39. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.

40. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. **13**(5): p. 323-30, 307-8.

41. Laskowski, R.A., et al., *Protein clefts in molecular recognition and function.* Protein Sci, 1996. **5**(12): p. 2438-52.

42. Carugo, O. and S. Pongor, *A normalized root-mean-square distance for comparing protein three-dimensional structures.* Protein Sci, 2001. **10**(7): p. 1470-3.

43. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.

44. Pintar, A., O. Carugo, and S. Pongor, *CX, an algorithm that identifies protruding atoms in proteins.* Bioinformatics, 2002. **18**(7): p. 980-4.

45. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.

46. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-9.

47. Hubbard, S., *NACCESS.* Department of Biochemistry and Molecular Biology, University College of London, 1993.

48. Colonna-Cesari, F. and C. Sander, *Excluded volume approximation to protein-solvent interaction. The solvent contact model.* Biophys J, 1990. **57**(5): p. 1103-7.

49. Fan, H. and A.E. Mark, *Refinement of homology-based protein structures by molecular dynamics simulation techniques.* Protein Sci, 2004. **13**(1): p. 211-20.

50. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Curr Opin Struct Biol, 2005. **15**(3): p. 285-9.

51. Rohl, C.A., et al., *Modeling structurally variable regions in homologous proteins with rosetta.* Proteins, 2004. **55**(3): p. 656-77.

CAPÍTOL 8.

CONCLUSIONS GENERALS

Tot i que les conclusions sobre els treballs realitzats s´han exposat al corresponents capítols, a continuació es resumeixen breument en cinc punts, corresponents als cinc capítols que formen el cos principal de la tesi.

Conclusió 1: pel que fa a la determinació de la família estructural en les prediccions *de novo* emprant xarxes neuronals i programes de comparació estructural, el rendiment del mètode és moderat; no obstant això el problema dista molt d´estar resolt, i no hem estat capaços d´aportar una solució definitiva.

Conclusió 2: en relació al mètode presentat per la extracció de les parts de millor qualitat de les prediccions *de novo* emprant comparacions estructurals amb proteïnes homòlogues, podem dir permet realment seleccionar aquelles parts de les prediccions millor predites. No obstant això mètode no és aplicable de forma directa, i depèn d´unes premisses com són: capacitat de determinar la família estructural de les prediccions, i presència d´homòlegs.

Conclusió 3: pel que fa al refinat dels models *de novo* emprant la informació sobre les parts de més qualitat, podem dir que el senzill protocol seguit permet aconseguir millores de qualitat, amb el que realment es pot considerar com una possibilitat alhora de refinar prediccions. El principal problema és que està lligat a les limitacions comentades en el punt anterior.

Conclusió 4: en relació a l´estudi de l´efecte de les mutacions puntuals en les cavitats del lisozim humà podem concloure que el principal efecte és a nivell de composició (per tant a nivell de característiques físico-químiques) i no tant a nivell geomètric.

Conclusió 5: pel que fa a l´estudi de la degradació del patró de cavitats en els models per homologia es pot concloure que com molt bé es pot pensar *a priori* a mesura que la qualitat dels models decreix, les distorsions de les cavitats són majors. De totes maneres hi ha dades interessants que apunten que en casos on el model per homologia és de mala qualitat és possible trobar cavitats ben modelades. Això obre una finestra al possible ús dels models per homologia en estudis de disseny de fàrmacs, fins i tot quan la seva qualitat global no és prou elevada.

LLISTA DE PUBLICACIONS

Preservation of protein clefts in comparative models.

Piedra D, Lois S, de la Cruz X.

BMC Struct Biol. 2008 Jan 16;8:2.

Involvement of chromatin and histone deacetylation in SV40 T antigen transcription regulation.

Valls E, Blanco-García N, Aquizu N, Piedra D, Estarás C, de la Cruz X, Martínez-Balbás MA.

Nucleic Acids Res. 2007;35(6):1958-68. Epub 2007 Mar 6.

A simple method for the identification of the correct parts in de novo protein structure predictions

Piedra D, de la Cruz, X.

Segona etapa de revisió a BMC Bioinformatics.

Protein function-related structural properties as components of organismal fitness

Calvo M, Piedra D, Lois S, Barbany M, de la Cruz, X.

Primera etapa de revisió a Journal of Evolutionary Biology.