Investigation of protein–ligand interactions using high-throughput all-atom molecular dynamics simulations

Ignasi Buch Mundó

TESI DOCTORAL UPF / ANY 2012

DIRECTOR DE LA TESI
Dr. Gianni De Fabritiis
Departament de Ciències Experimentals i de la Salut

UNIVERSITAT POMPEU FABRA

"May the GeForce be with you, always"
—Krunchin-Keith, *GPUGRID volunteer*

Acknowledgments

As so many other things in life, this thesis is an accident. I was set for a totally different path in the intersection between biology and computing when Gianni approached me in 2008 talking about molecular dynamics simulations on PlayStation's. I knew nothing about molecular simulations at that time, but it had to be certainly interesting if one could use a PlayStation for it. At home, walking around the kitchen table between stories of trains in life that come and go, I decided I was jumping on that one of uncertainly exciting destination. They were knowledgeable, passionate and innovative, the three basic ingredients for success. Time proved me right. In four years we have achieved a notable degree of visibility within our scientific niche, published several pieces of work in relevant journals and been awarded several times in national and international conferences. And this was just the beginning.

On the personal side, although it hasn't always been a pleasant journey—we're talking about a PhD thesis after all—, it hasn't been a solitary one either. I've had the pleasure and honor to share it with fellow travelers from around the globe. Some were here from beginning to end, some got on and off and others have just got on, and for better or for worse, they all contributed to who I am now. If I've ever smiled at you, you're one of them. Tank you very much.

Special mentions go to Gianni, my supervisor, and postdocs Toni and Kashif. I thank Gianni because he asked me to continue with the PhD after the masters on a day I was certainly convinced he was showing me the door. Unarmed, that struck me in a way that I had to stay. His ambitious ideas set the foundations of everything presented herein. I thank Toni and Kashif because if there have been two key influencing people at all levels to the successful completion of this work, it's them. Wizards of the arts of life and science, they always had great advice to give.

I want to also thank the crunchers, the enthusiastic GPUGRID volunteers who have been supporting us with their machines and donations. Thanks as well to the donors of La Marató de TV3, who have made my research economically possible. I hope that one day the technologies presented here can bring something back to society.

Finalment, als amics i als de casa us agraeixo el suport i la paciència d'aquests anys. Vista així acabada... potser no n'hi havia per tant, no? Moltes gràcies a tots!

Abstract

Investigation of protein–ligand interactions has been a long-standing application for molecular dynamics (MD) simulations given its importance to drug design. However, relevant timescales for biomolecular motions are orders of magnitude longer than the commonly accessed simulation times. Adequate sampling of biomolecular phase-space has therefore been a major challenge in computational modeling that has limited its applicability. The primary objective for this thesis has been the brute-force simulation of costly protein–ligand binding modeling experiments on a large computing infrastructure. We have built and developed GPUGRID: a peta-scale distributed computing infrastructure for high-throughput MD simulations. We have used GPUGRID for the calculation of protein–ligand binding free energies as well as for the reconstruction of binding processes through unguided ligand binding simulations. The promising results presented herein, may have set the grounds for future applications of high-throughput MD simulations to drug discovery programs.

Resum

La investigació d'interaccions proteïna–lligand és una important aplicació de les simulacions de dinàmica molecular (MD) donada la seva importància en el disseny de fàrmacs. Tanmateix, l'escala de temps rellevant per als moviments de biomolècules és molt superior als temps simulats habitualment. La simulació adequada de l'espai de fase és doncs una de les principals limitacions de l'MD. L'objectiu principal d'aquesta tesi ha estat la simulació per força bruta de costosos experiments de modelatge proteïna–lligand en una gran infraestructura computacional. Hem construït i desenvolupat GPUGRID: una infraestructura de computació distribuïda per a simulacions d'MD d'alt rendiment. Hem utilitzat GPUGRID pel càlcul d'energies lliures d'unió entre proteïna–lligand així com per a la reconstrucció de processos d'unió a partir de simulacions sense guiatge de lligand. Els prometedors resultats que es presenten, poden haver establert les bases de futures aplicacions de les simulacions d'MD d'alt rendiment en programes de descoberta de fàrmacs.

Preface

Thirty-five years ago McCammon, Gelin and Karplus presented the groundbreaking 9 ps molecular dynamics (MD) simulation *in vacuo* of bovine pancreatic trypsin inhibitor [1]. Since then, computer power has dramatically increased following the famous ''Moore's Law" bringing us today personal desktop computers that are millions of times faster that the devices used for MD in the 70s and 80s.

MD simulations are now used to study nearly every type of macromolecule—proteins, nucleic acids, lipids—of biological or medicinal interest. Simulations span wide spatial and temporal ranges and resolutions. In all-atom MD, thousands of individual atoms representing, for instance, all the atoms of a protein and surrounding water molecules, move in a series of femtosecond-long time steps. These movements repeated billions of times provide continuous atomic trajectories lasting as long as microseconds and, in very specific cases, milliseconds. Relevant biological motions such as protein folding, large conformational changes and protein–ligand interactions have timescales that are, at the very least, of hundreds of microseconds. Hence, to properly study these processes in atomic detail with MD simulations, tremendous amounts of computations will be required.

This thesis is focused on the particular problem of simulating protein–ligand binding processes with an eye for applications to drug discovery. Protein–ligand binding has been tackled since the near inception of MD always suffering from insufficient computational power and hence, sampling. This thesis has been developed around these particular issues. Our approach has been taking a big leap in accessible computer power and using it to address two of the most expensive modeling experiments in the field of protein–ligand interactions: binding affinity calculations and unbiased equilibrium-based ligand binding.

Specifically, we have built the GPUGRID project, a high-throughput computing platform for performing MD simulations on voluntarily-shared GPU-equipped desktop computers by thousands of people from around the world. We have been able to attract the attention of thousands of contributors who have allowed us to use their computers to perform some of the largest simulations ever reported. We have used GPUGRID to tackle the two aforementioned problems in protein–ligand modeling: the precise calculations of binding affinities for large and flexible ligands and quantitative reconstruction of ligand binding from unbi-

ased simulations. Both applications have been published in high impact journals, in particular, the quantitative reconstruction of binding for an enzyme–inhibitor system that became a hallmark study in the field. We are confident that the methods and applications developed have the potential to becoming useful tools in drug discovery in the near future.

Finally, the apparently spontaneous nature of the works presented in this thesis is, in fact, the reflection of a constant boundary-pushing exploration beyond state-of-the-art. We strongly believe that the thesis itself is a valuable outlook to what brute-force sampling approaches for protein–ligand binding can be capable of.

Contents

Chapter 1

INTRODUCTION

1.1 Protein–ligand interactions

1.1.1 Relevance and application

The formidable advances in protein sciences in recent years have highlighted the importance of protein–ligand and protein–protein interactions in biology. The majority of proteins in an eukaryotic cell are involved in complex formation at some point in the life of the cell and each protein has on average six to eight interacting partners [2]. Interactions can be classified on the basis of partner composition (homo/hetero-oligomers), independent occurrence of partners (obligate/non-obligate) and on stability of the complex [3, 4, 5], the latter defining interactions by their binding affinity on a continuum between transient and permanent.

At the structural level, protein interactions have been studied through crystallization of the complexes formed. The pioneering work on antigen–antibody and protease–inhibitor complexes provided insight into interacting interfaces and their properties [6, 7]. More recently, the structure of larger complexes that function as molecular machines has also been determined, shedding light into important cellular functions such as transcription [8], translation [9, 10], replication [11] or the cytoskeleton [12], to cite a few.

Understanding protein–ligand interactions is central to the design and discovery of new medicines too. Traditionally, drugs were discovered by trial and error. Drug discovery evolved to be increasingly deliberate and, with the advent

of structural biology, the rational design of inhibitors was made possible. Given the three-dimensional structure of a target enzyme for example, 'structure-based design' can be carried out, whereby an inhibitor is constructed to be complementary to the enzyme's active site [13, 14]. The minimum requirements are the target's structure and tools to build and examine how molecules fit into the active site. Additional insight provided by evaluating the molecular energetics of the binding process is, however, crucial to most current activities in structure-based design [15, 16]. This thesis deals with this particular requirement in the study of protein–ligand interaction through the development of methods to study and predict the energetic and kinetic binding features of inhibitors and naturally occurring ligands.

Molecular recognition models

From a mechanistic point of view, protein–ligand interactions can occur through three accepted models of molecular recognition, the 'lock-and-key' model [17], the 'induced fit' hypothesis [18] and 'conformational selection' [19]. In the 'lock-and-key' model, the conformations of the free and ligand-bound protein are similar, whereas 'induced fit' states that conformational differences between these two states are the result of the binding interaction driving the protein toward a new conformation that is more complementary to its binding partner and thus energetically more favorable. The 'conformational selection' model proposes that, given a conformational heterogeneity, weakly populated (higher energy conformations) are responsible for recognizing and binding to partners with subsequent population shift toward these conformers [20, 19]. Heated discussions have been going on for years now [21] on which flexibility acknowledging model, induced fit or conformational selection, described more accurately molecular recognition. The conclusion is that both models do in fact coexist. In this direction, the most accepted molecular recognition pathway model seems to be that where kinetic rate constants would dictate which pathway is followed by the system [19, 22].

In a recent study on a large scale analysis of 2090 unique unbound to bound transitions from over 12,000 solved structures, Orozco and co-workers [23] showed that two-thirds of the analyzed complexes did not suffer significant structural changes and could thus fit the lock-and-key model. Among the remaining ones, they reported one-third of the proteins exploring the bound conformation in the unbound state, which would fit into the conformational selection model, and

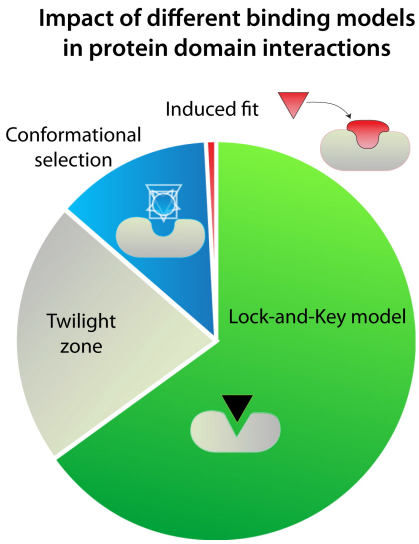

Figure 1.1: After analyzing 2090 unbound to bound structural transitions out of 12,000 protein structures, Stein et al. [23] found that 65% of the proteins did not undergo major conformational changes upon binding, a 13% explored their bound conformations in their unbound state and only 2% required a external energy push to reach their bound conformation. Figure adapted from Stein et al. [23].

only a very few transitions required breakage of thermodynamic barriers to bind, a definition of the induced-fit model (Figure 1.1). Altogether, this flexibility in protein–ligand interactions imposes a major challenge for drug discovery practitioners since the comforting idea that there is one ligand perfectly adapted for one static protein structure is outdated. Although some efforts have already been using searches through ensembles of conformations to find one matching and accommodating the ligand of interest with fair amounts of success [24, 25, 26], it is still an inexact approximation of biological reality. Part of the work presented in this thesis is in fact around the role of conformational flexibility for specific protein–ligand interactions [27]. The future of drug design will thus require tools to deal with flexible biological molecules, with the great potential of opening up the possibility to explore broader chemical spaces than the currently available [28, 29, 30].

1.1.2 Biophysical aspects of interactions

Thermodynamics and kinetics of binding

The reversible binding of a ligand (L) to a protein (P) can be written as:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL \tag{1.1}$$

where, assuming a first-order kinetic model, k_{on} ($\mathrm{M}^{-1}\mathrm{s}^{-1}$) and k_{off} (s^{-1}) are the association and dissociation rate constants respectively. The equilibrium constant can be derived as

$$K_{eq} = \frac{[PL]}{[P][L]} = \frac{k_{on}}{k_{off}}, \tag{1.2}$$

where $[P]$ and $[L]$ are the protein and ligand concentrations. Binding affinities are expressed in terms of the equilibrium dissociation constant (K_D) which is

$$K_D = \frac{1}{K_{eq}}. \tag{1.3}$$

A more general and comparable measure for binding affinity is the concentration-independent standard Gibbs binding free energy (ΔG°_{bind}) obtained form the equilibrium constant K_{eq} through the well-known formula [31, 32]

$$\Delta G^{\circ}_{bind} = -k_B T \ln C^{\circ} K_{eq}, \tag{1.4}$$

where k_B is the Boltzmann constant, T the temperature and C° the standard state concentration (1 M). The more negative the value of ΔG°_{bind}, the more favorable binding is. The change in free energy itself is composed of enthalpic (ΔH) and entropic (ΔS) changes. ΔH is, effectively, the heat given out or taken up upon making and breaking interactions, and ΔS represents the energetic consequences of changes to the degrees of freedom within the system, where

$$\Delta G_{bind} = \Delta H - T\Delta S. \tag{1.5}$$

A variety of physical phenomena are thought to contribute to the binding free energy of an interaction, including those that are considered to make a largely enthalpic contribution, for example, van der Waals interactions, hydrogen bonding and electrostatic complementarity, and those considered to be dominated by entropy, for example, changes in configurational disorder and in the solvation of

hydrophobic/lipophilic groups upon formation of the complex [33, 34]. All these structural determinants of protein–ligand interactions have been studied for many years with the purpose of inferring energetic structure–activity relationships that can be applied to the design and discovery of strong inhibitors [35, 36, 37].

Binding kinetics is increasingly receiving a lot of attention too in the characterization of protein–ligand binding in the context of drug design. Some reviews for instance, have highlighted the fact that there exist stronger correlations between *in vivo* activities of some drugs with their residence times that with their binding affinities to their targets [38, 39]. An analysis from Swinney [40, 41] revealed that for drugs approved by the FDA between 2001 and 2004, 34% had non-equilibrium kinetics and 31% were known to induce conformational changes in proteins. Conversely, a study of cyclooxygenase inhibitors suggested that rapid dissociation rates are a means of minimizing mechanism-based side effects [42]. Hence it is reasonable to conclude that greater consideration of optimal kinetics at the time of clinical candidate selection will lead to reduced attrition during development and that it will be possible to differentiate future drugs on the basis of their kinetics. Slow off-rates are desirable in the absence of mechanism based toxicity to ensure maximum target engagement and enhanced specificity resulting in greater safety margins and reduced adverse events. Rapid off-rates are desirable where there is mechanism-based toxicity as a means of minimizing these effects. In summary, identification of kinetic mechanisms in biomolecular recognition and their optimal combinations [43], opens up a new era for medicinal chemistry by incorporating kinetic structure–activity relationships to drug discovery processes.

Experimental methods for binding affinity, kinetics and structure determination

Accurate measurement of binding affinities and kinetics as well as production of high-resolution structures is of paramount importance to the study of protein–ligand interactions. For binding affinity measurements, isothermal titration calorimetry [37] is often the method of choice due to its high precision and ability to specifically determine enthalpic/entropic contributions. Surface plasmon resonance is also used in binding affinity measurements but through determination of association and dissociation rates, the binding kinetics [44]. On structure (and dynamics) determination there are X-ray crystallography and Nuclear magnetic resonance (NMR) [45, 46]. Although NMR is widely used to

study the dynamics of proteins, they are most widely known for their ability to solve the protein structure at an atomic level. As opposed to being a substitutive, computer-based molecular simulation techniques like the ones presented in this thesis, have its best use in the interpretation of the results of experimental techniques by providing atomic-scale views of the phenomena under study [47].

Isothermal titration calorimetry (ITC) is a physical technique used to determine the thermodynamic parameters of interactions in solution. It is most often used to study the binding of small molecules to larger macromolecules since ΔG°, ΔH° and $T\Delta S^\circ$ can be accurately determined from a single experiment [48]. In an ITC experiment, the incremental heats of reaction are measured as one component is titrated into the other and ΔH° and ΔG are determined by nonlinear fitting of the resulting titration curve [49]. The entropy change associated with interaction can then be determined from equation (1.5). Although a variety of other techniques can be used to accurately determine the affinities (K_D or ΔG°) of protein–ligand interactions, ITC experiments produce much more accurate sets of thermodynamic parameters for protein–ligand interactions than have previously been available, providing greater reliability of the data used to assess the relationship between structure and thermodynamics [37].

Surface plasmon resonance (SPR) spectroscopy is an electromagnetic wave resonance-based technique widely used to monitor a broad range of analyte-surface binding interactions including the adsorption of small molecules [44], ligand-receptor binding [50], protein adsorption on self-assembled monolayers [51], antibody-antigen binding [52], DNA and RNA hybridization [53] and protein-DNA interactions [54]. The sensing mechanism of SPR spectroscopy is based on the measurement of small changes in refractive index that occur in response to analyte binding at or near the surface of a noble metal (Au, Ag, Cu) thin film [55]. SPR has the advantage of being label-free [56]; capable of probing complex mixtures, such as clinical material, without prior purification [55]; and benefits from the availability of commercial instrumentation with advanced microfluidic sample handling [57, 58]. As a biosensor technology it can be used both qualitatively and quantitatively to monitor protein–ligand interactions. In a qualitative screening mode, receptor binders and non-binders can be identified. In a quantitative high-resolution mode, precise kinetic and affinity parameters can be obtained across a wide dynamic range. [50, 59, 60, 61]

X-ray crystallography allows the determination of the arrangement of atoms within a crystal, by striking a crystal with a beam of X-rays that spreads

into many specific directions. From the angles and intensities of the diffracted beam, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal. From this electron density, the mean positions of the atoms in the crystal can be determined, as well as their chemical bonds, their disorder and various other information [45]. X-ray crystallography can also be used to study dynamics, especially of slow timescales. Given that for high-resolution X-ray crystallography, homogeneous crystals are needed, in order to observe protein substates one has to trap them through biochemical 'tricks', or synchronize a reaction across an entire crystal [62, 63, 64].

Nuclear magnetic resonance spectroscopy is based on the property of many elements to have a nuclear magnetic moment—of particular importance in biological macromolecules are the stable isotopes H^1, C^{13} or N^{15}. When placed into a static magnetic field B, the different nuclear spin states of these nuclei become quantized with energies proportional to their projection onto B (the so-called Zeeman Splitting). The energy difference depends on the type of nucleus, is proportional to field strength of the static magnet, and is dependent on the chemical environment of the nucleus. This energy difference corresponds to electromagnetic radiation. The transition between these states can be induced by irradiation with a radio-frequency field with characteristic frequencies for each type of nucleus and its chemical environment. The frequency of the NMR signal is extremely sensitive towards changes in covalent bonds such as neighboring groups and also to noncovalent bonding as found in biomolecular interaction. Furthermore, transfer of magnetization through bonds or through space results in a characteristic change of the shape and size of the NMR signal and reflects, for example, the bond angle in the case of scalar coupling or spatial distance in the case of dipolar coupling. All these phenomena are exploited in several applications aimed at resolving the three dimensional structure of proteins or characterizing protein–ligand interactions among others [64, 46].

1.2 Molecular dynamics modeling

Molecular dynamics (MD) is a computational technique to simulate the motions of a system of particles. The essential elements for an MD simulation are a knowledge of the interaction potential of the particles, from which the forces can be calculated, and of the equations of motion governing the dynamics of the particles [65]. MD simulations model biomolecular systems as point-like

masses moving with the action of classical forces. A simulation begins with an initial set of atomic coordinates and velocities. Coordinates can be obtained from X-ray crystallographic or NMR structure data, or alternatively, by homology model building (based on the structure of a homologous protein) [66]. Velocities, obtained by solving the classical Newtonian equations of motion derived from Newton's law $\vec{F_i} = m_i\vec{a_i}$, are updated at each time step (Δt) millions of times so that biologically relevant events can be observed [67]. The sum of forces ($\vec{F_i}$) are derived from a set of interaction potentials between atoms defined in the 'force field' parameters file (see Figure 1.2).

MD currently faces several important challenges in modeling biomolecular function related to the computational cost associated with the sampling of biologically relevant timescales, the accuracy of the simulations and the prediction of statistical quantities comparable to experiments. The following sections briefly cover these issues as well as introduce the main sampling and analysis methods used throughout the development of this thesis.

1.2.1 The sampling and force field issues

For MD simulations to reliably reproduce, guide and help explain experiments, three things are required: adequate sampling of the relevant biomolecular motions, force fields of sufficient accuracy and correct representation of the experimental conditions. Although force fields are the most commonly blamed issue when performing MD simulations, in a way, adequate sampling may be the weakest point. Until sampling is adequate, equilibrium properties computed from a simulation remain biased by the system's starting state and no meaningful comparison with experiment is possible. On the other hand, robust although disagreeing results are still possible with inadequate force fields or poor representation of experimental conditions [68, 69]. Issues of adequate sampling and force field accuracy are briefly covered in the following paragraphs.

Addressing the sampling problem is in fact the primary focus of this thesis. One limitation of current MD simulations of biomolecules is that many important biomolecular motions take place with characteristic timescales much longer than typical simulation timescales [64]. As a consequence, experiments are reported with clearly insufficient simulation times and often not enough attention is put into testing the adequacy of sampling. For example, ligand binding modes are slow change, presenting problems for binding mode prediction [68, 70]; protein conformational changes even at the single sidechain level can be slow, affect-

ing the quality of the computed binding free energies [71, 72]; slow motion of waters into and out of binding sites can hurt convergence and thus apparent accuracy [73], and unsampled protein conformational changes can also introduce errors [71]. Still, despite efforts to make it a standard practice for reproducibility [74], reliable indication of convergence is hard to be found [75, 76, 77]. In a recent perspective review, Mobley [69] suggests that the vast majority of the "accuracy" problems in the literature about protein–ligand binding modeling can be traced back to specific sampling problems. This suggests that sampling may be a leading cause of error and that these are real problems of precision stemming from the mismatch between available simulation and biomolecular-motion timescales. Even more, if and only if adequate sampling is achieved, we can quantitatively assess the accuracy of a particular force field, identify deficiencies, and improve it [78]. Fortunately, recent improvements available computational powers are pushing forward simulation timescales. These include the building of specialized supercomputing architectures [79], porting MD software to consumer-market multiprocessor devices like GPUs [80] and exploiting volunteered distributed domestic desktop computer networks as exposed in following sections [81, 82].

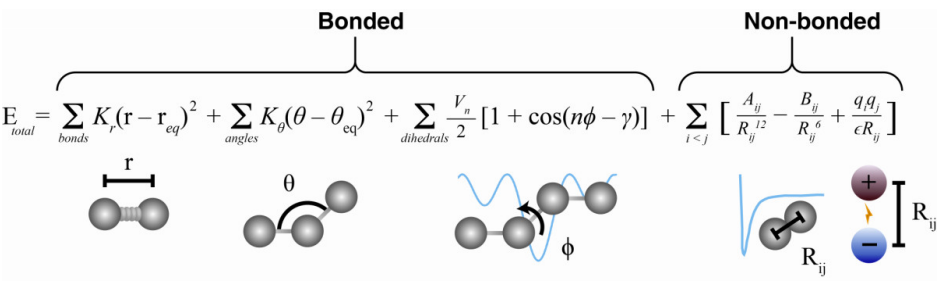

Figure 1.2: An example of a potential used to approximate the inter-atomic forces that govern molecular movement. The equation is divided into terms treating interactions between atoms that are chemically bonded and into those treating interactions between atoms that are at long distances. Figure adapted from Durrant and McCammon [83]

The accuracy of MD computations is also a concern, and is ultimately determined by the underlying molecular mechanics force field. Figure 1.2 shows an example of an equation used to approximate the atomic forces that govern molecular movement. While the mathematical functional forms of many of the available force fields are quite similar, they differ in the parameters that describe

the various energetic components and in the methods employed to obtain these parameters. These parameters and hence the force fields that stem from them, are in constant refinement. As an example, in the last few years a sequence of studies have revised and modified torsion potentials associated with a few important dihedral angles. Simmerling and co-workers [84] modified the backbone potential in the original Amber ff99 force field by fitting to additional quantum-level data deriving the improved Amber ff99SB force field. Best and co-workers followed up on this work by modifying the backbone potential in ff99SB and ff03 to obtain a better energetic balance between helix and coil conformations, thus producing the ff99SB* and ff03* force fields [85]. Lindorff-Larssen and co-workers modified the side-chain torsion potential for four amino acid types in ff99SB to produce the ff99SB-ILDN force field [86] and more recently, they combined it with ff99SB* to produce ff99SB*-ILDN and also changed parameters associated with both the backbone and certain side chains in a CHARMM force field to produce CHARMM22* [87].

In a hallmark force field validation study, scientists at D.E. Shaw Research recently presented a systematic comparison of a number of force fields for all-atom simulations in explicit solvent that combined several common tests for force field evaluation performed over unprecedented simulation lengths [88]. They evaluated the recently developed force fields mentioned above as well as their original versions. They tested a total of nine different force field versions including OPLS-AA [89, 90], Amber ff99SB-ILDN [84, 86], Amber ff99SB*-ILDN [84, 86], Amber ff03 [91], Amber ff03* [91, 85], CHARMM22 [92], CHARMM27 (CHARMM22 with CMAP correction) [92, 93] and CHARMM22* [87]. Their final conclusion based on reviewing force field performance on matching experimental NMR results of folded proteins, temperature-dependent structural propensities in short peptides and folding of α-helical and β-sheet proteins, determined that newer versions CHARMM22* and Amber ff99SB*-ILDN are the top performing force fields for these type of tests [88]. Although the study has not been evaluated the impact of force field on more complex molecular systems i.e. membrane-embedded proteins, it is certainly a relevant and necessary contribution to the field.

All of the aforementioned force fields however, suffer from transferability issues. They are designed to model proteins, nucleotides and some lipids, but are not readily transferable to any chemical compound, a serious problem for modeling protein–ligand interactions. To overcome this limitation, ‘generalized’ force

fields such as GAFF [94] with AM1-BCC [95] fixed partial charge model, have been developed and are widely used to parametrize, for instance, drug molecules.

Nonetheless, neither the specific nor the generalized force fields take into account yet several important determinants in protein–ligand interactions which limits the accuracy of the representations. Some examples are the effect of induced electronic polarizability, that when incorporated has been shown to achieve high accuracy results in binding free energy calculations [96], changes in protonation states upon binding [33] and the existence of tautomers [97]. For accuracy, these shall be incorporated into routine parametrization of protein–ligand complexes in the near future.

1.2.2 Binding free energy calculations

In previous sections we have introduced the importance of binding free energy (ΔG) as measure of the strength of an interaction. In the context of structure-based drug design, a great focus is put on the accurate computations of binding free energies to evaluate potential drug candidates in early-stage drug discovery [15]. Several computer methods have been developed to approach the calculation of affinities in a trade off between speed versus physical accuracy. The fastest and less physically accurate methods are grouped around the concepts of molecular docking [98, 99, 100, 101, 102] and approximate free energy methods such as the linear interaction energy (LIE) methods [103, 104] or the molecular mechanics Poisson-Boltzmann/Generalized-Born solvent accessible surface area (MM-PBSA/GBSA) methods [105, 106, 107, 108, 109], in which solvent and protein motions are taken into account with fewer approximations.

On the other hand we have the slow but accurate, true free energy methods that use conformational sampling to generate thermodynamic averages, to compute either the free energy difference between the bound and unbound state through decoupling the interactions between the ligand and its receptor (alchemical double decoupling schemes) giving a non-physical pathway, or to compute differences as well, but most importantly, absolute binding free energies by displacing the ligand along a physical pathway of binding (pathway-based methods). Free energy perturbation (FEP) [110, 111, 112] and thermodynamic integration (TI) [113, 114, 115] are alchemical double decoupling methods for binding free energy calculations traditionally employed for—but not limited to—calculating relative binding free energies between related protein–ligand combinations, being able to calculate absolute binding free energies [116]. The lat-

ter however, incurring in a much larger computational cost. Methods involving the biased sampling along a set of pre-selected reaction coordinates that follow physically meaningful binding pathways include, among others, metadynamics [117, 118, 119], steered MD [120, 121, 122] and umbrella sampling [123, 124, 71] which is later described in more detail.

The result of a computational free energy calculation can be only as accurate as the force field used to generate the ensemble. In general, the best performing protocols show a mean error of around 1 kcal/mol, but there are much larger deviations expected depending on the nature and size of the compound [125, 126, 127]. However, as discussed earlier in the text it is well appreciated that the main problem with free energy simulations is their difficulty to converge. Mobley et al. [128] reviewed the contribution to binding affinities of the motions, ensembles, alternative conformers, entropies and forces 'unseen' in single molecular structure studies. Also, in methods requiring in principle less conformational sampling there seems to be a large dependence of the binding free energies computed to the ligand poses used [129, 130, 72].

To estimate the uncertainty on a computed free energy change, block averaging techniques are commonly used. To do so, an entire trajectory is divided into blocks and free energy changes are computed with the data available in each block and the standard deviation and mean of the free energy changes provides an estimate of the stability of the computed free energy changes [81]. However, a more expensive but arguably better way to obtain good estimates of the statistical error is to repeat each free energy simulation independently [121]. The free energy change and associated error can then be estimated from the mean and standard deviation of the independent realizations [71, 121].

Comparison with experimental data

A word of caution is necessary when attempting to compare predicted binding free energies to measured binding affinities. Simulations usually imply idealized conditions such as pure water, which rarely reflect the conditions in which binding affinities are measured, or because the protein used in the assay differs somewhat from the protein structure used for the predictions [131]. A binding assay is often setup to measure binding affinities within a limited range, ITC for example, works at the sub-millimolar to nanomolar range for direct measurement of binding constants and between nanomolar to picomolar for competitive assays; binders out of these ranges, may be under or overestimated [37]. Also,

often binding free energies come from titration experiments which may suffer from poor parameter fitting of the kinetic curves [132].

A key issue in the comparison of binding free energy results is the comparability of this via the definition of a standard state [32, 133]. Binding constants are defined in terms of ratios between reactants and products (see Eq.(1.2)) which presents a problem when their numbers are not the same and the binding constant is not dimensionless in equation (1.4). In other words, in order to calculate a free energy from an equilibrium constant whether it is via theory, simulation or experiment, we must define a standard state so that we can make meaningful comparisons between them. In a recent work, General [32] presented a detailed and unified explanation to convert a binding free energy from an arbitrary state to some given standard state. In this regard, this thesis has adopted the standard binding free energy expressions derived by Doudou et al. [134] for the computation of pathway-based free energies over one-dimensional or three-dimensional coordinates of reaction [81, 71, 135].

Potential of mean force

The potential of mean force (PMF) $W(z)$ along some generalized coordinate $z(\vec{x})$ (Figure 1.3a), is a key concept in statistical mechanics. It is the product of physical pathway-based free energy sampling methods and it is defined as the negative logarithm of the probability of being at a given value (state) of a specified reaction coordinate

$$W(z) = -k_b T \ln p(z), \tag{1.6}$$

where k_b is the Boltzmann constant and T the temperature and $p(z)$ the probability of being at a specific value in z. This reaction coordinate may be an angle, a distance or a more complicated function of the Cartesian coordinates of the system. Generally, any conformational equilibrium properties can be expressed in terms of the function $W(z)$. For these reasons the PMF is a central quantity in computational studies of macromolecular systems [123]. However, it is often impractical to compute $W(z)$ directly from MD simulations. The presence of large barriers in z may not allow accurate sampling of the configurational space within a finite computational time. An example is the case for binding-pathway free energy calculations. Binding of macromolecules often occurs at the microsecond-millisecond scale [64] which is highly costly and therefore hardly

achievable even in high-performance/high-throughput sampling scenarios [27]. This is why, to avoid such difficulties, special sampling techniques have been developed over the years to calculate the PMF from MD trajectories efficiently. An example of enhanced sampling method or technique is the aforementioned umbrella sampling [136], which has been implemented in this thesis for computation of binding free energy calculations of protein–ligand systems [81, 71].

1.2.3 Umbrella sampling

Umbrella sampling is a physical pathway-based (PMF-based) sampling technique [136]. In umbrella sampling, the system of interest is simulated in the presence of an artificial biasing window potential, $v(z)$, introduced to enhance the sampling in the neighborhood of the chosen value z. The biased simulations will be generated using the potential energy $U + v(z)$, where U represents the total energy of the unbiased system. The biasing potential will typically confine the variations of z within a small interval around some prescribed value (the window center), helping to achieve a more efficient configurational space sampling in this region. An often used choice of biasing potential is an harmonic function of the form $v_i(z) = 0.5k(z - z_i)^2$, centered on successive values of z_i. To obtain the PMF over the whole range of interest of z one will need to perform a number of biased window simulations, each biasing the configurational sampling around a different region of z (Figure 1.3b). Ultimately, the results of the various windows are unbiased and then recombined together to obtain the final estimate $W(z)$ [136, 123].

The weighted histogram analysis method (WHAM) reconstructs the PMF from biased umbrella sampling data [137, 138]. The basic idea of the method consists in constructing an optimal estimate of the unbiased distribution function as a weighted sum over the data extracted from all the simulations and determining the functional form of the weight factors that minimizes the statistical error.

The precise estimation of free energies using umbrella sampling depends first and foremost, on the choice of reaction coordinate; a badly chosen coordinate of reaction will result in poor free energy estimations [139], a general problem for physical pathway-based free energy methods [117, 134]. An other factor affecting the precision of umbrella sampling is for example, the degree of overlap between the windows which is a result of the compromise between number of windows along a reaction coordinate and the width of the umbrella

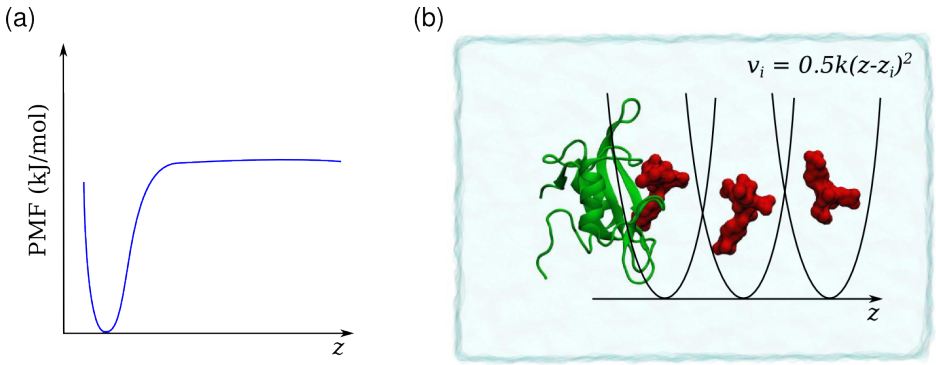

Figure 1.3: Representation of a schematic pathway-based protein–ligand interaction. (a) Barrier-less one-dimensional potential of mean force (PMF) of the distance z between two interacting partners. (b) Simplification of an umbrella sampling biasing scheme with three windows to reconstruct the PMF.

potentials controlled via the force constant k of the harmonic function [123]. Ultimately though, the convergence will be dominated by the choice of reaction coordinate and the relaxation times of the internal degrees of freedom of the system. In publication 3.2 of this thesis, we present a study on the convergence and accuracy successes of an umbrella sampling-based protocol where we explore several combinations of parameters and starting configuration sets to enhance binding free energy estimations [71].

Notable work in the application of umbrella sampling and WHAM to compute binding free energies on one-dimensional PMFs has been performed by Roux and co-workers [124, 140, 141]. They have developed and applied extensively an approach based on applying multiple restraints to the ligand, determining the radial one-dimensional PMF, and then removing the restraints. The standard free energy of binding was then obtained using a system-specific derivation which made it impractical for a general case [124]. On the other hand, Henchmann and co-workers [134], using a similar methodology than Roux's, presented a much simplified application of the umbrella sampling for the computation of the standard free energy of binding using a one-dimensional PMF. Works on which publications 3.1 and 3.2 of this thesis are based.

1.2.4 Markov State Modeling

A radically different view in the sampling and analysis of biomolecular dynamics is brought by Markov State Models (MSMs), a tool based on finite-state transition networks [142, 143, 144, 145, 146]. MSMs are probabilistic models that can be built from molecular dynamics data to approximate long-time statistical dynamics of molecules. Figure 1.4 shows a sample representation of a two-state Markov model.

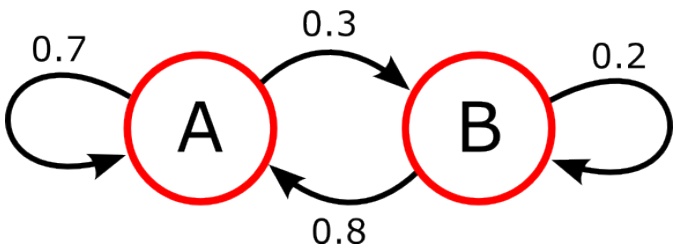

Figure 1.4: Schematic representation of a two-state Markov model. According to the markovian property, the probabilities of transition between states (indicated with numbers), only depend on the current state of the process. For example, at each discrete time interval, if the system is in state A, it has 30% probability of transiting to B and 70% probability of staying in state A.

To date, MSMs have been used to computationally model the kinetics and thermodynamics of complex molecular conformational transitions such as protein folding [142, 147, 148, 149, 150] or protein–ligand binding [151, 152, 135]. They have also been applied to the modeling of experimental outcomes; an example is the work by Kusch et al. [153] where they kinetically quantified all ligand binding steps and closed-open isomerizations of the intermediate states of the activation mechanism of homotetrameric HCN2 channels from confocal patch-clamp fluorometry data.

MSMs have represented a paradigm shift in how one uses simulations [144, 145, 146]. Traditionally, MD studies have relied on straightforward simulations and analysis of a few rare events based on a 'look and see' strategy. Although visually appealing, these analyzes do not provide sufficient statistical relevance of the observations and therefore may be highly misleading in reporting on important events altogether. MSMs on the contrary, abandon this single view of trajectories to substitute them by an ensemble view of the dynamics and hence

can be used to resolve measurable statistical properties of the ensemble: time-dependent averages of spectroscopically observable quantities, statistical probabilities quantifying which conformational substates are populated at certain times and probabilities of how many trajectories follow similar pathways [146].

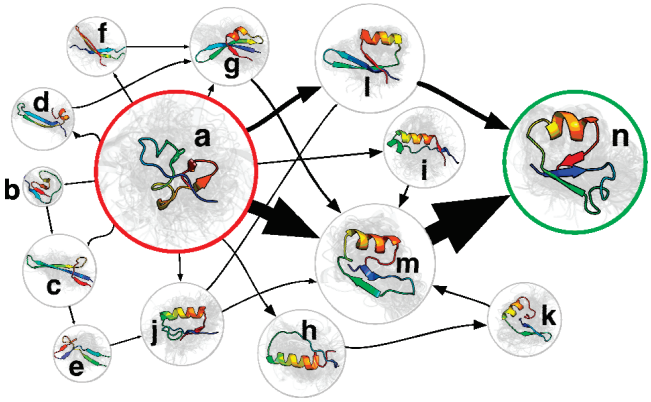

Figure 1.5: Transition network for the folding of NTL9(1-39), adapted from Voelz et al. [148]. The underlying Markov State Model contained 2000 states and was built from 10000 individual implicit solvent MD simulations. The visual size of each state is proportional to its free energy, and arrow size is proportional to the interstate flux.

MSMs also does away with the frequent approach of projecting the dynamics onto one or two user-defined coordinates of reaction by means of enhancing sampling of *a priori* important reaction coordinates (i.e. umbrella sampling), or by means of analyzing bias-free trajectories. When projecting trajectory data on few coordinates of reaction, one hopes that such projection will capture the slow kinetics of the process under study. However, these projection techniques often disguise the true and complex nature of kinetics by artificially aggregating kinetically distinct structures and hiding barriers, with the potential danger of creating distorted and overly simplistic pictures of the kinetics [154, 144, 146].

An interesting feature of MSMs is the possibility of reconstructing long time-scale dynamics from from many short MD trajectories. A feature that makes MSMs a very suitable modeling tool for highly parallel infrastructures as it has been done for this thesis. Since MSMs are constructed only from the conditional transition probabilities, short trajectories need only to be long enough as the local equilibration time within the states and occasionally undergo transition between

states. Pioneering works in protein folding by Noé and Pande have demonstrated this feature of MSMs. Noé and co-workers [147] presented the reconstruction of the full equilibrium ensemble of folding pathways of a PinWW domain from MD simulations in explicit solvent. From a total of 180 individual trajectories of 115 ns each they recovered the ensemble of pathways, the slowest timescale of which was two orders of magnitude larger than the simulated trajectories. In a similar fashion, Pande and co-workers [148] in a highly parallel effort employing the Folding@Home distributed computing platform [155], reconstructed the folding of 39-residue protein NTL9(1-39) via MSM (shown in Figure 1.5), with an experimental folding time of 1.5 ms. Using up to 10,000 parallel implicit solvent MD simulations of around 10 to 15 ns each, they predicted the millisecond timescale of the folding of the protein under study. For protein–ligand binding, Silva et al. [152] and us in Publication 3.4 showed how from short MD trajectories too, binding events could be recovered with decent accuracies for unguided binding simulations.

Given the relative novelty of the application of MSMs to MD, the following paragraphs are provided as a brief reference for model building, validation and interpretation:

Discretization of the state space

The state space of the system has to be discretized into a set of $S = \{1, ..., m\}$ conformational states. Each data point (or microstate) is assigned into a macrostate for example by geometrical proximity. The number of macrostates of our model will greatly vary between studies and may depend on the specific system/mechanism under study, the microstate clustering method used—if any—and of course on the level of detail desired. The discretization step is a fundamental part of MSM building. Prinz et al. [146] showed how the quality of the Markov model strongly depends on how well the discretization approximates the slow processes of the system.

Construction of the transition probability matrix

An $m \times m$ transition probability matrix $T(\tau)$ is then constructed, where each element T_{ij} measures the probability of going from state i to state j within time τ, by $T_{ij} = c_{ij} / \sum_k c_{ik}$. Here, c_{ik} counts the number of times the trajectory was in i at time t and in j at time τ later. Although this expression provides the most likely

transition matrix, the full probability distribution of $T(\tau)$, given c_{ij}, must be considered when statistical uncertainties of $T(\tau)$ and properties computed from it are desired [147, 146]. Additionally, $T(\tau)$ is required to be ergodic, i.e., any state can be reached from any other state within a finite time. Then, $T(\tau)$ has a single eigenvector with eigenvalue 1 that, when normalized, the stationary probability π is obtained. For equilibrium MD, π is be the equilibrium distribution and altogether should hold the detailed balance condition: $\pi_i T_{ij} = \pi_j T_{ji}$ [147, 146].

Assessment of markovianity

A markovian system is a memoryless system, the state transition probabilities must only depend on the current state and not on the past. The most frequent cause for non-markovianity is the presence of state-internal barriers [144]. All models will be Markovian for long enough lag times, τ, but to maximize the time resolution of a model, shorter lag times are always desired. A straightforward test for markovianity is the convergence study of implied timescales for the main transition modes with increasing lag times. This will determine too the minimal lag time needed for the model to remain Markovian. Other methods include Chapman-Kolmogorov tests for example, which check markovianity on a particular state decomposition by assessing if conformations within a state do kinetically interconvert on timescales faster than the lag time and only make transitions to other states on slower timescales [145, 147, 146].

Interpretation of the model

In principle, any property that can be calculated from simulation data can also be obtained from the MSM. The first property that we will be interested in analyzing is the equilibrium distribution of our system. The equilibrium distribution can be directly obtained from the elements in the first left eigenvector of $T(\tau)$ as mentioned above.

Another property of interest of special relevance in the modeling of protein–ligand binding is the mean first passage time (MFPT), defined as the mean time f_i it takes to reach a given metastable state m for the first time when starting from another state i [144]. In binding, with a binary view of the states and assuming first-order kinetics, one can compute the MFPTs for the on and off reactions from which derive k_{on} and k_{off} respectively from the expressions $k_{on} = C^{-1}\mathrm{MFPT}_{on}^{-1}$

and $k_{off} = \mathrm{MFPT}_{off}^{-1}$ [142, 156, 135]. Where C is the ligand concentration and k_{on} is measured in $\mathrm{M}^{-1}\mathrm{s}^{-1}$ and k_{off} measured in s^{-1}.

Finally, one can also compute the net fluxes through the reaction pathways. This features allows the classification of reaction pathways in terms of best, next best, etc and although it has not been used for this thesis, one can already find applications in protein–ligand binding of multistep complex systems [151], protein folding [147, 148] or protein conformational changes (S.K. Sadiq unpublished data on HIV-1 protease).

1.3 High-throughput MD simulations

Computing power has traditionally been and still is a limiting factor when performing molecular simulations. There is a constant need for faster computing resources to reduce time-to-answer and increase sampling times in computational experiments. Computational speed of processors is, on average, doubled every 18-24 months, a trend known as the Moore's Law [157]. Traditional approaches aim at using faster processors or parallelization of computer programs to run on many processors that are available either on commodity clusters or large supercomputing facilities [158]. An outstanding contribution to this matter in the last 10 years has been that of DE Shaw Research, a private research laboratory who has built a specialized ultra-high-performance computer for molecular simulations, Anton [79]. Scientists at DE Shaw Research have performed all-atom MD simulations in explicit solvent models of up to the millisecond time scale for single trajectories [67]. Unfortunately, a single Anton computer has a production cost of the scale of several million dollars, certainly prohibitive to large-scale production.

1.3.1 Accelerated processors for molecular dynamics simulations

Fortunately, high-performance cost-effective solutions are possible. Recent hardware development led by the gaming industry needs, has led to a breakthrough in accelerated processors: general-purpose computing architectures, the graphical processing units (GPUs). GPU boards are hardware accessories targeted to the market of personal desktop computers to off-load the display of graphics from the computer's central processing unit (CPU). They are equipped with hundreds (i.e 512 in an NVIDIA GeForce 580) of small processors that are able to perform

independent computations on independent data. Together with the architecture itself, GPU manufacturers they also offered accessible programming interfaces that allow for a wider spread adaptation and creation of new computer codes to exploit the hardware's processing capabilities.

Benefiting from the momentum that GPUs have created, MD has made it into the accelerated processors realm [158]. Pioneered by ACEMD software [80, 159], today the majority of codes in the field (i.e. NAMD [160], AMBER [161], GROMACS [162]) do offer the ability to run on GPUs with varying degrees of performance as represented in Figure 1.6. As of the time of writing of this thesis, current state of ACEMD software running on high-end GPUs on a single workstation it is possible to break the symbolic barrier of 100 ns a day for a 23,000 atoms system with an explicit solvent model (timestep of 4 fs, PME for long-range electrostatic interactions and a cutoff of 9 Å for non-bonded interactions).

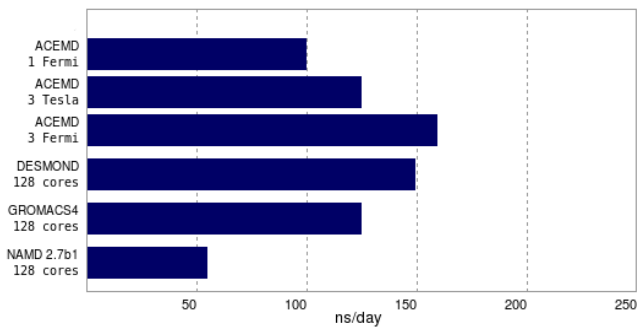

Figure 1.6: Comparative performance of ACEMD versus other common MD codes. Fermi is a GTX580 GPU. Tesla is a C2050 with ECC off. CUDA3.1 and ACEMD ver 2011. DHFR, dihydrofolate reductase solvated in water, 23558 atoms, periodic boundary conditions, 9 Å cutoff, PME long range electrostatic $64 \times 64 \times 64$, hydrogen mass repartitioning, rigid bonds, Langevin thermostat, time step 4 fs. Time step 2 fs NAMD [160], 2.5 fs DESMOND [163] and 4 fs GROMACS [162].

1.3.2 Volunteer distributed computing: the GPUGRID project

Another form of computing infrastructure are computing grids, distributed computers loosely connected often through the Internet or local networks. In fact,

the majority of today's computing power is distributed among a billion domestic computers in the world most of which have access to the Internet. They are an under-exploited computational asset. The possibility of using domestic distributed computing power was embraced by scientists more than a decade ago. But it is not until 2002 that the Berkeley Open Infrastructure for Network Computing (BOINC) [164] middleware software is born to serve the needs of pioneering project SETI@home [165]. Volunteer distributed computing became then accessible to the scientific community who started exploiting domestic computers from volunteers for all kinds of projects. From radio-telescope signal processing to climate change prediction modeling, projects started to incorporate to exploit the world's largest computing facility. On aggregate, and as of March 2012, BOINC has 300,000 active participants who contribute 470,000 hosts in total, an average computing capacity of 6 PetaFLOPS. Currently, the largest conventional supercomputer on the planet is the 'K computer' in Japan, which provides a theoretical peak performance of 11 PetaFLOPS [166].

From a technical point of view, to be amenable to public computing, computational scientific tasks must be divisible into independent pieces whose ratio of computation to data transfer is high. While this consideration is maintained, distributed computing is a real alternative to classical dedicated supercomputing facilities. A key factor of such computing system is engagement of public individuals, the volunteers. Also referred to as participants or users, volunteers donate time from their personal computing resources to scientific computing driven by two main motivations: their interest in the underlying science and public acknowledgment. Unlike in other forms of social scientific computing like Foldit [167] were users are actively solving specific scientific puzzles, in the distributed computing community, the volunteers select the scientific projects in which they want to contribute based on their personal interests, and these could be several simultaneously. In exchange, the project acknowledges the volunteers' contribution via a credit or points system. After a piece of work is finished by the volunteer and returned to the project servers, the project gives a certain amount of credits to the volunteer. On top of this, volunteers self-organize themselves in virtual communities, exchanging information through message boards and sharing their gained credits in rankings of various natures. At this point, their altruistic contribution to science is transformed into a game, the goal of which is to accumulate as many credits as possible.

The GPUGRID project

GPUGRID [168] is a BOINC-based volunteer distributed computing project that exploits the power of consumer-level GPUs to perform high-throughput and high-performance molecular dynamics simulations by running the ACEMD software since 2008 (Figure1.7). Formerly known as PS3GRID and exploiting the power of volunteered PlayStation 3, GPUGRID has, as of March 2012, an active volunteer base of 2600 users and 2800 host computers (see Figure1.8 for growth pattern) each one having one or more GPU cards attached to the project that make up a total of 3500 GPUs which represents a theoretical peak performance of 1.6 PetaFLOPS. Figure 1.9 shows the distribution of GPU peak performance in active hosts and the distribution of GPUs installed per host. With regards to MD data production, GPUGRID outputs a daily average of 22 μs approximately of an equivalent system of 50,000 atoms.

With such throughput it becomes crucial to automate as much as possible the procedures of interaction with the server. From the scientist point of view, the main handicap in using a BOINC-based distributed computing environment is the actual submission of the computations to the grid. This operation was traditionally done manually and explicitly by logging into the web server and executing the task-submission applications. Giorgino et al. [169] developed RBoinc, an interface with mechanisms to submit and manage large-scale distributed computations from individual workstations turning distributed grids into cost-effective virtual resources.

As mentioned earlier, the main motivations for the public to participate as volunteers in distributed computing projects are the interest in the underlying science and the acknowledgment or public recognition. In GPUGRID we have understood these motivations and have actively worked towards providing better participation experiences. For that, we have taken several actions to address these issues. We have created and maintained project-specific web pages, audiovisual resources and forum discussions to promote their research to the contributors, fundamentally a lay audience. Such efforts have been widely appreciated and have often generated fruitful discussions with the volunteers about the nature and impact of the projects in which they participated. Regarding the public acknowledgment aspect of volunteers motivations, BOINC is already designed to assign volunteers a number of points or credits per completed task. Such credits are used to rank the users by their contribution to the project within a number of communities. Nevertheless, a number of BOINC-related projects had cre-

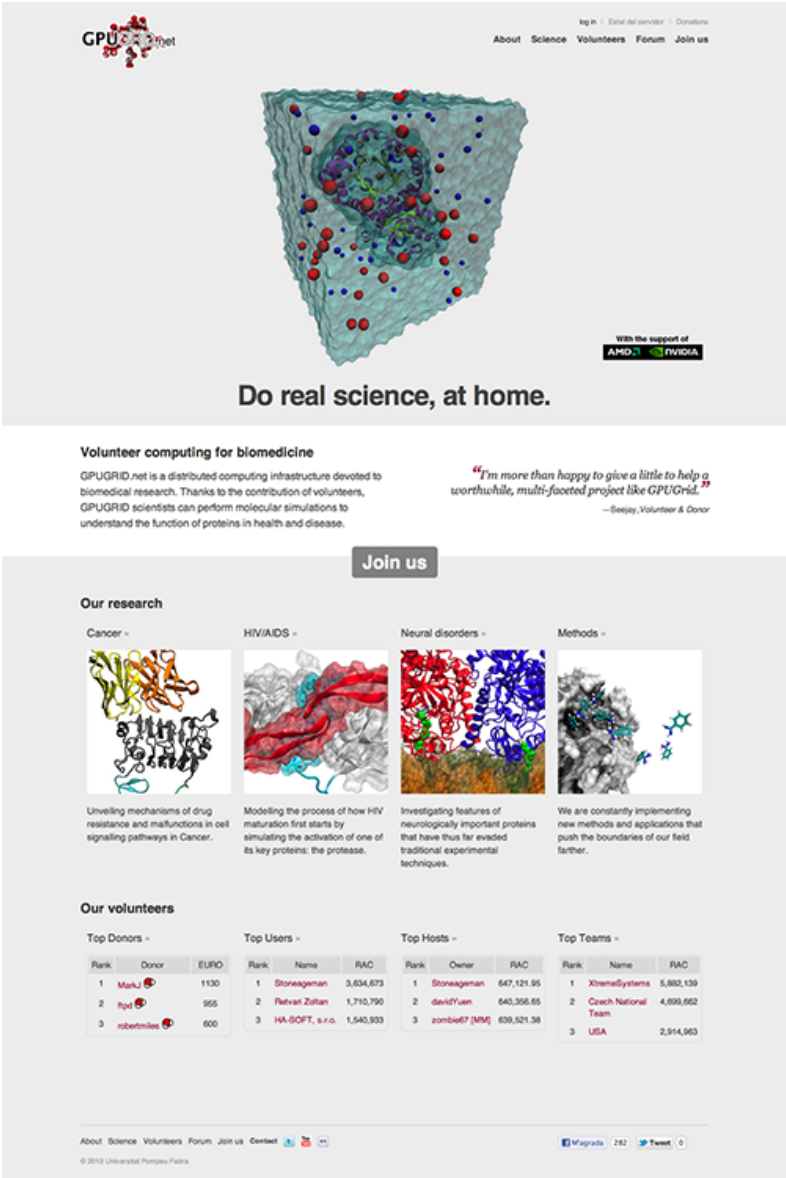

Figure 1.7: GPUGRID website as of January 2012. The website is the main vehicle of communication between the scientists and the volunteers. It features information on the scientific projects being executed on the grid, the message boards or forums for discussions on technical and scientific aspects of the project and leader boards or rankings displaying per-user contribution on the project.

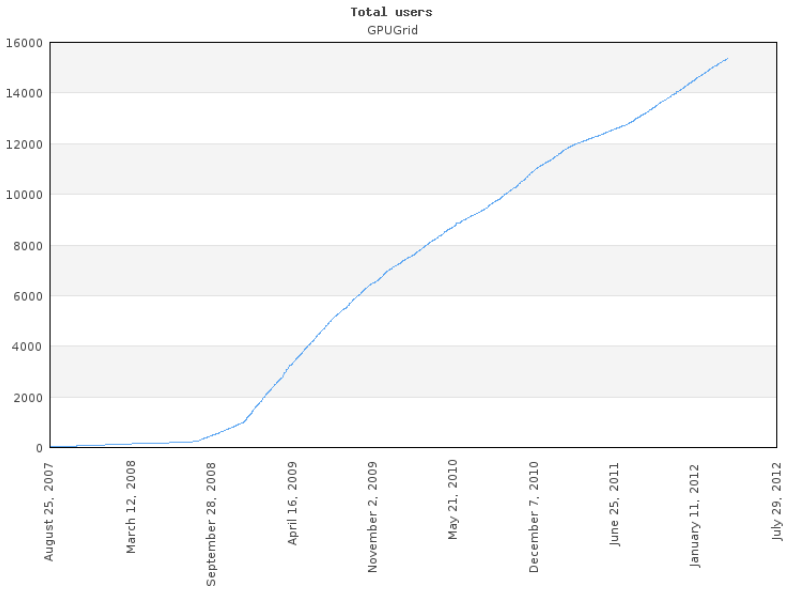

Figure 1.8: Evolution of GPUGRID user base since 2007, when the project was born as PS3GRID. Total number of users is, at the time of writing of this thesis, is over 15000 of which only 17% are active. The effective number of contributing users is around 2600. Plot and data obtained from AllProjectStats [170]

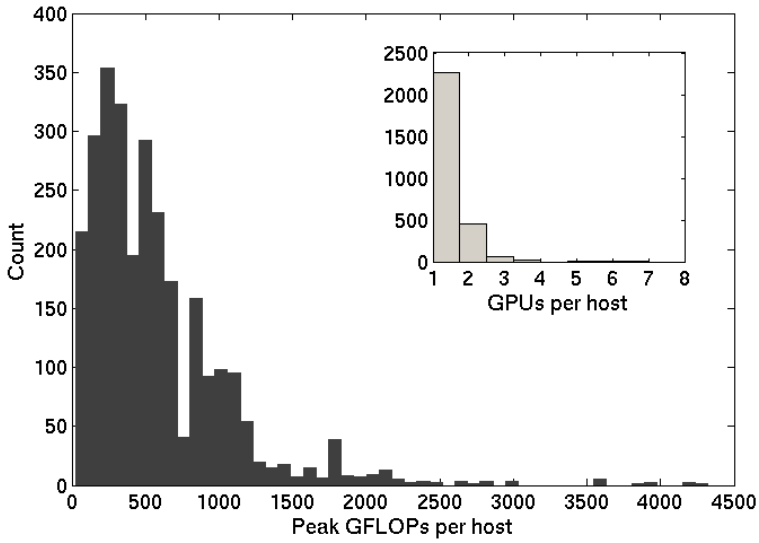

Figure 1.9: Distribution of GPU peak performance installed in active hosts of GPUGRID (main plot) and GPUs installed per host (inset). During the period considered, 2800 active hosts (3500 GPUs) provided a theoretical processing power (to be adjusted by the fraction of resource shared) of 1.6 PetaFLOPS.

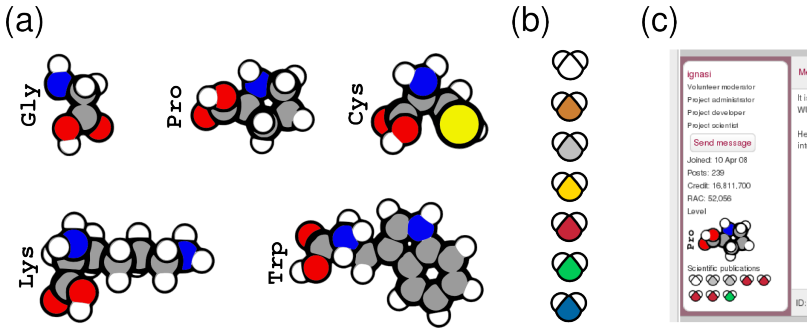

Figure 1.10: Badges given to users to acknowledge their contribution to the project. (a) Badges representing the twenty encoded amino acids by the universal genetic code are used to visually indicate the accumulated number of credits of a volunteer. A table of equivalences is used between the number of credits and amino acid badge ordered by molar mass, being glycine to easiest to achieve (500,000 credits) and tryptophan the hardest (10 billion credits). (b) Water molecule badges are used to indicate the user relative contribution to individual scientific publications. Seven levels were named after precious metals and stones ordered by value: paper, bronze, silver, gold, ruby, emerald and sapphire. (c) Badges are displayed together with the volunteer's information on their user profiles.

ated in the past visual badges or prizes, a form of recognition that related to the amount of credits obtained by displaying a small graphical icon. Examples are found in projects such as World Community Grid [171], Primegrid [172] or Yoyo@home [173]. GPUGRID has not been an exception to this trend. With the help of a representative of the volunteer community, we have designed one of the most innovative badge systems. GPUGRID now grants its participants for the absolute contribution measured by the total number of credits obtained and for the contribution to the actual publications stemming from GPUGRID data. This last form of acknowledgment has been widely appreciated by the community as it condenses and makes evident the user contribution to the scientific knowledge produced and presented with publications. Not surprisingly, careful attention and acknowledgment of the volunteers resulted in spectacular increases of participation specially since the implementation of the badge system. Examples of the badges can be seen in Figure 1.10. The amino acids, ordered by molar mass, are used to represent total contributed computations in GPUGRID. With a fixed table of equivalences, whenever a user reaches certain threshold, a heavier amino acid badge is assigned to her profile. The second system, that acknowledges the user for their contribution in scientific publications, assigns cumulative water molecules colored after precious metals and stones. Each water molecule links to the corresponding publication and scientific explanation page. Such engagement of participants via game-like frameworks through token-per-task assignments is known as 'gamification' [174] and is an emerging trend in Social Marketing for mass-consumer industries [175].

1.4 Macromolecular systems studied

In the course of this thesis we have focused our efforts on three molecular systems; two well-studied systems for method development and one as an application of the methods. The systems for method development have been an Src-homolgy 2 domain with a phosphorylated ligand and trypsin with an inhibitor. We have applied some of the methodology to a more complex protein–protein interaction system, the Epidermal growth factor receptor with an antibody and with a ligand. The following sections briefly describe the systems and some of their most important features.

1.4.1 Src-homology 2 domain

We have used a SH2 domain-phosphopeptide complex as a test case to develop binding affinity calculation protocols [81, 71] as well as to experiment with highly flexible unguided binding simulations of peptides for the computation of binding kinetics and binding pathway reconstruction [27].

Src-homology 2 domain (SH2) domains were first identified as non-catalytic modules conserved among members of the src family of cytoplasmic protein-tyrosine kinases [176]. They have since been found in many other proteins that are involved in intracellular signal transduction [177]. SH2 domains can be found in proteins that possess enzymatic activity (for example, kinases or phosphatases). Alternatively, they can be present in adaptor proteins that lack any catalytic activity (for example, Grb2, which contains one SH2 and two SH3 domains [178]). SH2 domains bind phosphotyrosine-containing peptides of selected sequences with high affinity [179] as represented in Figure 1.11. The recognition of phosphotyrosine-containing motifs in activated cell surface receptors by the appropriate SH2 domains is an important step in the intracellular signal transduction process. Due to their essential role in the signal transduction process and their selectivity towards phosphotyrosyl peptide sequences, SH2 domains are potential targets for therapeutic intervention. Specially in cancer treatments where many signal transduction routes appear altered [180, 181]. For a review see Pawson and Gish [177].

1.4.2 Trypsin

We have performed binding simulations of bovine *beta*-trypsin and benzamidine to reconstruct complete binding processes, recovering affinity, kinetics and binding pathway [135].

Trypsin is serine protease enzyme that hydrolyzes other proteins and polypeptides. Serine proteases, among which we find chemotrypsin and elastase, receive the name from a very well conserved serine residue that performs a crucial role in the catalytic mechanism of action. Serine proteases are found in the digestive system of many vertebrates, where they perform their protein hydrolysis function to permit the absorption of amino acids through the lining of the small intestine. Serine proteases are produced in the pancreas in the form of zymogen proenzymes, inactive forms of the enzyme, that are initially activated by enteropeptidases and later activated through autocatalysis. All serine

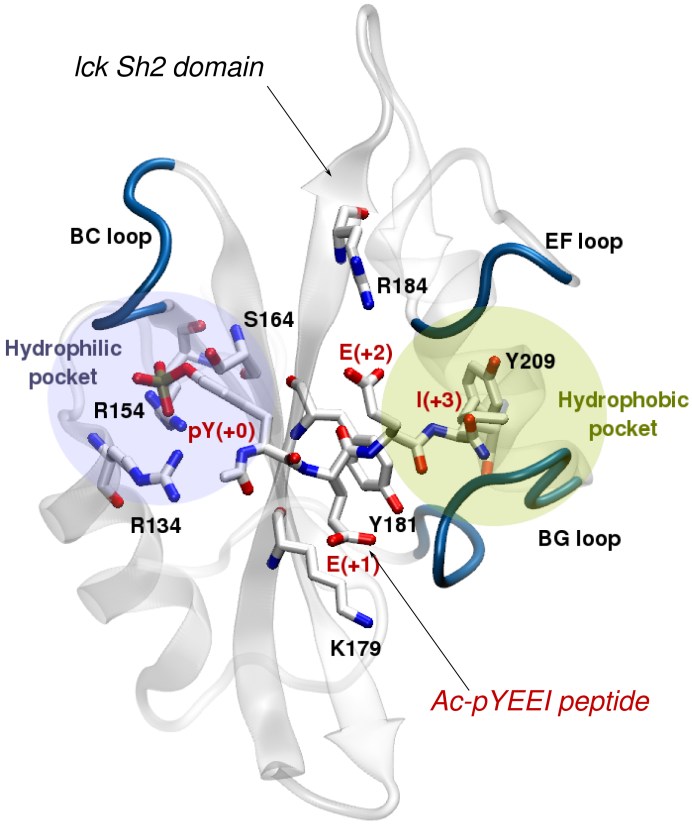

Figure 1.11: 1LKK PDB crystal structure of the SH2 domain of human p56lck in complex with the short phosphotyrosyl peptide Ac-pTyr-Glu-Glu-Ile (pYEEI peptide) [180]. The protein's secondary structure (transparent) and relevant loops are highlighted (blue). The pYEEI peptide (sticks) plugs into two pockets: a hydrophobic one shown on the left, "proximal", which buries phosphotyrosine pY(+0), and a hydrophobic one shown on the right, "distal", accommodating I(+3). Significant residues forming native contacts between the protein and the peptide are labeled in black and red, respectively. Secondary structure elements are named according to Eck et al. [182]. Figure adapted from Giorgino et al. [27].

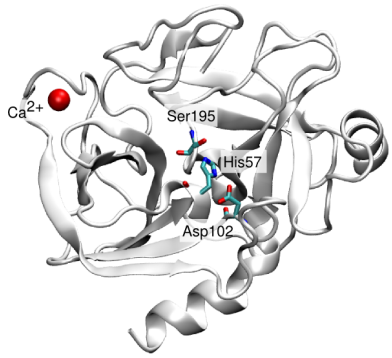

Figure 1.12: Crystal structure of bovine β-Trypsin with calcium ion bound in a regulatory site [188]. The protein's solvent accessible surface area is shown in white except for the catalytic residues His57, Asp102 and Ser195, rendered in licorice representation.

proteases have a preferential point of cut at the carboxylic side of amino acids. Trypsin, for example, cuts the carboxylic side of basic residues such as lysine or arginine while chemotrypsin preferentially cuts after an hydrophobic residue, i.e. phenylananine [45]. This catalytic mechanism is performed by the residue triad formed between His57, Asp102 and Ser195 as shown in Figure 1.12. The mechanism of action has been known for many years now [183] although new families of enzymes utilizing the mechanism are being discovered, in which the nucleophile-base-acid pattern is generally conserved, but the individual components can vary [184]. Trypsin was one of the first protein whose structures were solved by x-ray crystallography [185, 186]. Trypsin has often been crystallized with small inhibitors such as benzamidine [187, 188], which later also resulted in numerous methodological works on benzamidine derivatives that have aimed at understanding the thermodynamic contributions of substituents to binding [36] as well as being a test-bed of computational methods for binding free energy [117, 134].

1.4.3 Epidermal growth factor receptor

We have applied our previously developed binding affinity calculations protocol [71] to predict the binding affinities for drug cetuximab and ligand EGF for the wild-type and the mutant receptor to determine the impact of a mutation on

complex formation.

Ligand-induced signaling from receptor tyrosine kinases (RTKs) of the epidermal growth factor receptor (EGFR) family (also known as ErbB or HER) regulates many cellular processes, including proliferation, cell motility, and differentiation [189] (see Figure 1.13). Perturbations in these cellular signals can lead to malignant transformation, and the correlation between EGFR and cancer has been firmly established [190]. Deregulation of EGFR can arise from its over-expression [191], mutation/truncation of the receptor [192], or activation by aberrant autocrine growth factor loops [193]. EGFR has been implicated in the development of a wide range of epithelial cancers, including those of the breast, colon, head and neck, kidney, lung, pancreas, and prostate. In these settings, deregulation of EGFR correlates with decreased disease-free and overall survival [194, 195, 196, 197]. EGFR is currently being targeted in anticancer treatment via monoclonal Antibodies such as cetuximab [52] and panitumumab [198]. Recently however, a missense S468R mutation on the ligand and drug binding EGFR domain III has been described to differentially affect the treatment of colorectal cancers with cetuximab and panitumumab [199]. Upon emergence of this mutation, malignant cells develop resistance to cetuximab but not to panitumumab despite sharing binding site.

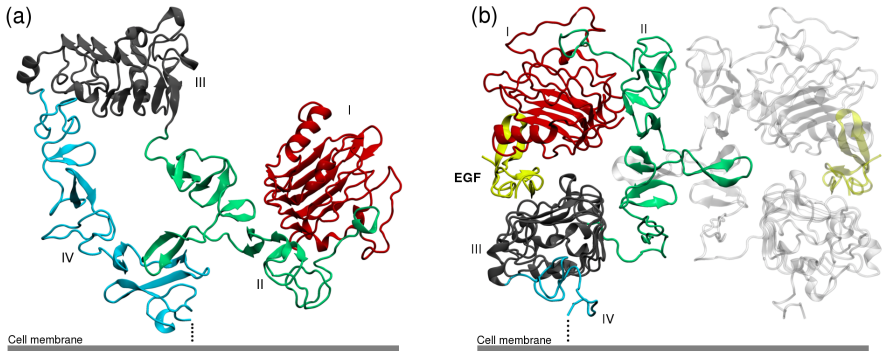

Figure 1.13: Crystal structure of the single chain EGFR extracellular domain in (a) the untethered conformation [52] and (b) the tethered ligand-bound dimerized conformation [61]. The extracellular single chain domain is composed by four different sub-domains namely I (shown in red), II (shown in green), III (shown in grey) and IV (shown in cyan). Upon ligand binding, EGFR adopts the tethered conformation that allows it to dimerize with another receptor single chain. With dimerization, the intracellular domains of EGFR cross-phosphorylate each other which recruits further signaling proteins that propagate the signal transduction inside the cell.

Chapter 2

OBJECTIVES

The main objective for this PhD project has been the development and application of a powerful computing infrastructure to studying protein–ligand binding, typically limited by sampling capacities. From this, we derived several sub-objectives that were gradually established and addressed throughout the thesis and can be stated as follows:

2.1 Setup and development GPUGRID for high-throughput molecular dynamics simulations

Relevant biomolecular motions like binding or conformational changes, have characteristic timescales beyond the microsecond. Simulating and, more importantly, calculating thermodynamic and kinetic properties for these motions requires large amounts of computation. The applicability of MD is therefore limited by computing capacity.

We have specifically addressed this limitation by setting up GPUGRID, a volunteer distributed computing infrastructure made of GPU-equipped personal computers. On top of GPUGRID we have implemented protocols to routinely perform high-throughput MD simulations of binding free energy calculations as well as discovery and quantification of rare molecular events. Publication 3.1 addresses the implementation of MD protocols on distributed computing network GPUGRID.

2.2 Implementation and application a one-dimensional potential of mean force-based method for binding free energy calculations

Calculating protein–ligand absolute binding free energies is a long-standing goal for molecular modeling. Its main application is the discovery of ligands that bind targeted proteins with high affinities. Among the numerous methods available at various physical accuracies, computing binding affinities using explicit solvent physical pathway-based interaction representations is the most accurate strategy but it is often computationally prohibitive and requires of expert human intervention.

We have addressed these limitations through the implementation and optimization of a one-dimensional potential of mean force protocol for the precise and accurate calculations of absolute binding free energies on GPUGRID. Publication 3.1 presents the first implementation of the protocol for the SH2–pYEEI system, further optimized for performance and precision in publication 3.2 on the same system. Finally, in publication 3.3 we present the application to EGFR–cetuximab and EGFR–EGF systems to evaluate comparatively the impact of a mutation on drug and ligand binding in the context of cancer treatment.

2.3 Implementation and application of unbiased sampling methods for complete binding process reconstruction

The grand challenge in the study of protein–ligand interactions is the direct observation and quantification of unbiased equilibrium-based ligand binding at atomic resolution, something which has remained at a prohibitive computational cost until now.

We have employed GPUGRID to perform unbiased protein–ligand binding simulations. These unbiased simulations have unveiled complex processes for the interactions such as the existence of non-native metastable states or the relationship between ligand binding and receptor flexibility. We have also applied MSMs for the analysis of the unbiased data to calculate binding affinities, kinetics and pathways for the interactions. Publication 3.4 shows a complete application of unbiased binding of trypsin–benzamidine for full quantitative recon-

struction using MSMs. Publication 3.5 shows unbiased binding of SH2–pYEEI discussing the roles of conformational flexibility.

Chapter 3

PUBLICATIONS

3.1 High-throughput all-atom molecular dynamics simulations using distributed computing

Buch I., Harvey M.J., Giorgino T., Anderson D.P. and De Fabritiis G., *Journal of Chemical Information and Modeling* 50, 397 (2010)

Summary

In this work we reviewed the innovations in accelerating molecular dynamics on graphics processing units (GPUs), and we described GPUGRID, a volunteer computing project that uses the GPU resources of non-dedicated desktop and workstation computers. We also demonstrated the capability of simulating thousands of all-atom molecular trajectories generated at an average of 20 ns per day each (for systems of 30,000-80,000 atoms) at the time. We then applied the resources of GPUGRID for binding free energy calculations of the Src SH2 domain–pYEEI ligand system, a rather complex system due to its size and flexibility. We applied a non-optimized version of an umbrella sampling-based potential of mean force (PMF) protocol and obtained a standard free energy of binding of -8.7 ± 0.4 kcal/mol within 0.7 kcal/mol from experimental results. The work proved that GPUGRID was a robust system for high-throughput binding affinity calculations.

Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G. High-throughput all-atom molecular dynamics simulations using distributed computing. J Chem Inf Model. 2010 Mar 22;50 (3):397-403.

3.2 Optimized potential of mean force calculations of standard binding free energy

Buch I., Sadiq S.K. and De Fabritiis G., *Journal of Chemical Theory and Computation* 7, 1765–1772 (2011)

Summary

Following from our previous work on implementing binding free energy calculations on GPUGRID, here we presented an optimized version of the one-dimensional potential of mean force method based on ensemble umbrella sampling simulations. The tests on the SH2 domain–pYEEI ligand resulted in an accurate and converged binding free energy of -9.0 ± 0.5 kcal/mol (compared to an experimental value of -8.0 ± 0.1 kcal/mol). We found that a minimum of 300 ns of sampling was required for every prediction. We described how convergence was obtained by using an ensemble of simulations per window, each starting from different initial conformations, and by optimizing window-width, orthogonal restraints, reaction coordinate harmonic potentials, and window-sample time. We also found that the use of uncorrelated initial conformations in neighboring windows was important for correctly sampling conformational transitions from the unbound to bound states that affected significantly the precision of the calculations. This methodology thus provides a general recipe for reproducible and practical computations of binding free energies for a class of semi-rigid protein–ligand systems, within the limit of the accuracy of the force field used.

Buch I , Sadiq S.K, De Fabritiis,G. Optimized potential of mean force calculations of standard binding free energy. *J Chem. Theory Comput.* 2011; 7(6): 1765 -1772.

Buch I , Sadiq S.K, De Fabritiis,G. Optimized potential of mean force calculations of standard binding free energy. Supporting information. *J Chem Theory Comput.* 2011; 7(6): 1765 -1772.

3.3 Computational modeling of cetuximab resistance to EGFR S468R mutant in colorectal cancer treatment

Buch I. and De Fabritiis G., *Unpublished manuscript* (2012)

Summary

Here we applied the optimized protocol for binding affinity calculations to provide a molecular structure-based explanation of the recently described acquired mutation in EGFR that causes resistance to treatment with cetuximab of colorectal cancer. By inspecting the bound structures of cetuximab, alternative antibody necitumumab and three EGFR ligands, we determine the putative impact of the mutation in their bindings. To confirm the structural analysis, we performed binding free energy calculations using the previously employed protocols based on one-dimensional potential of mean forces sampled by umbrella sampling, of cetuximab and EGF to both wild type and S468R mutant variants of EGFR. We predict a loss of affinity for cetuximab of at least 1 kcal/mol and an increase in affinity for EGF of about 1.1 kcal/mol. Although in need of experimental validation, we can propose a model in which cetuximab would be outcompeted by endogenous ligand EGF that would make treatment against this mutant variant ineffective. All in all, this work serves both as an application for our previously implemented protocol for binding free energy calculations as well as an example of the applicability of molecular modeling to rationalize drug usage in the context of personalized medicine.

Computational modeling of cetuximab resistance to EGFR S468R mutant in colorectal cancer treatment

The recently described S468R mutation in the extracellular domain III of the Epidermal Growth Factor Receptor (EGFR) causes resistance to cetuximab in colorectal cancer treatment. We performed a molecular structure-based assessment study to discuss the putative impact of the mutation on the binding of cetuximab, necitumumab and EGFR ligands EGF, TGFα and HRGa. We also apply molecular modeling techniques to calculate binding free energies for cetuximab and EGF to wild type and S468R mutant EGFR to specifically quantify the impact of the mutation to drug and ligand binding. Our results suggest that the S468R mutation may have a particularly deleterious effect on the efficacy of cetuximab in blocking receptor activation, due to a loss in cetuximab affinity and a gain in EGF affinity for EGFR. According to our predictions, mild alterations in opposite directions of binding affinities may be the reason to the resistance to cetuximab by S468R EGFR. This work provides an interesting example of application of high-throughput all-atom molecular dynamics simulations for an accurate prediction of resistance to monoclonal antibody-based therapy in the context of personalized medicine.

I. INTRODUCTION

Colorectal cancer is the third-leading cause of cancer-related deaths worldwide, with over 600,000 deaths occurring worldwide each year[1]. Recently, a role has been established for the epidermal growth factor receptor (EGFR) signal transduction pathway in the development of a subset of epithelial tumors[2]. EGFR is involved in multiple cellular proliferation processes, including growth, differentiation, migration, and apoptosis. EGFR over-expression has been shown to predict tumor progression[3] in colorectal cancer and is over-expressed in 25-77% of these tumors. EGFR is often associated with a worse prognosis[4].

In recent years, many EGFR-targeted agents have been developed. The two agents that have demonstrated the best responses are two monoclonal antibodies directed against EGFR: cetuximab and panitumumab[5] (known as anti-EGFR therapy or EGFR inhibitors) and compete against endogenous EGFR ligands like EGF for binding site as well as blocking receptor dimerization[6] (see Figure 1). These antibodies have presented high response rates when administered with chemotherapy. Cetuximab is a chimeric IgG1 anti-EGFR monoclonal antibody that has demonstrated anti-tumor activity in patients with colorectal cancer[7]. Cetuximab has a murine structural component which is a potential source of toxicity and immunogenicity[8]. Due to this, there has been a considerable amount of research aimed at eliminating this toxicity. As a result, a new agent was developed: panitumumab, a fully human IgG2 monoclonal antibody that is highly selective for EGFR[5]. Both cetuximab and panitumumab are considered fully equivalent in the treatment of colorectal cancer and therefore it is assumed that both share the same epitope[9,10]. However, a new missense mutation has been identified in the extracellular domain III of EGFR, S492R (S468R according to residue numbering in FabC225/EGFR crystal structure by Li et al.[6] and used herein). The mutation has been identified as the cause for acquired resistance to clinical treatment of colorectal cancer with cetuximab but, surprisingly, not with panitumumab; which has led to the conclusion that the two must recognize different epitopes of EGFR[10]. Unlike for cetuximab[6] though, there is no publicly available crystal structure for panitumumab that can aid to a proper structure-based analysis of the phenomenon.

Several mutations in domain III in EGFR have been previously reported in the literature to help understand the role of epitopic residues to the binding of cetuximab. Specially deleterious have been mutations Q408M in combination to H409E which caused a 150-fold decrease in FabC225 binding[6] as well as Q384A that in combination with the previous two caused a 380-fold decrease in binding[6]. Same sites but other mutations Q408A/H409A also produce a 10-fold decrease in FabC225 binding to sEGFR or 50-fold decrease again if combined with Q384A[11]. Milder decreases of 1.5-fold have been seen for K443A and S468I/N473A[11]. The study by Montagut et al.[10] is the first example of a missense mutation of the target of an antibody being the direct cause of resistance to that therapeutic antibody. Understanding the mechanisms of drug resistance can clearly lead to the development of more effective targeted therapies, new therapeutic combinations or both[9].

In this work, we perform a molecular structure-based assessment study to discuss the putative impact of the S468R EGFR mutation on binding of cetuximab, necitumumab and EGFR ligands EGF, TGFα and HRGa. We also apply molecular modeling techniques to calculate binding free energies for cetuximab and EGF to wild type and S468R mutant EGFR to specifically quantify the impact of the mutation to drug and ligand binding. Calculating binding free energies using molecular dynamics simulations (MD) is a widely explored topic in the field of computational biophysics and several methodologies have been successfully developed in recent years[12–15]. Here, we apply previously described protocols for high-throughput binding free energy calculations[16] of rather large and semi-rigid protein-protein complexes to compute the binding free energies of cetuximab and EGF to EGFR.

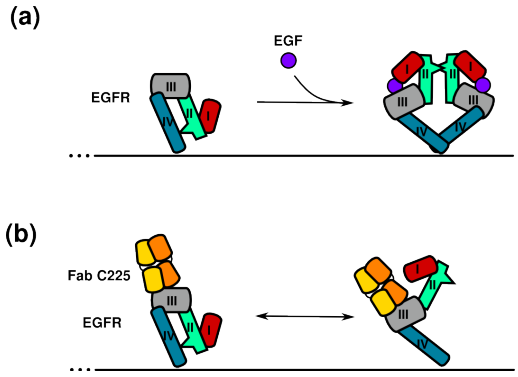

FIG. 1: (a) Sketch model of EGFR receptor dimerization induced by EGF binding to domains III and I. Dimerization is required for intracellular signal transduction. (b) Cetuximab (FabC225) as well as panitumumab, binds to domain III of EGFR blocking ligand binding and preventing the receptor from adopting an extended conformation that would permit dimerization.

II. METHODS

System preparation Structures of the bound FabC225 (cetuximab) to wild type sEGFR (PDB:1YY9)[6], bound Fab11F8 (necitumumab) to wild type sEGFR (PDB:3B2V)[11], bound EGF to wild type sEGFR (PDB:1IVO)[17], bound TGFa to wild type EGFR (PDB:1MOX)[18] and Neuregulin-1/HRGa (PDB:1HAF)[19] were obtained from the Protein Data Bank[20]. Cetuximab–EGFR, EGF–EGFR were used for MD simulations. From here on we will refer to these systems as 'cetuximab system' and 'EGF system' . Only interacting domains of the complexes were included in the simulations. Given the large size of cetuximab's Fab fragment, only the Fv domains of the antibody (residues 1-120 for the heavy chain and 1-108 for the light chain) and domain III of EGFR (residues 310-501) were used. In the case of EGF, the entire ligand and domain III of EGFR (residues 310-501) were used. All systems were parametrized using the CHARMM27 force field and solvated in a TIP3P water[21] boxes with ionic strengths of 0.15 M. The cetuximab system was solvated in a $80.0 \times 75.3 \times 137.0$ Å^3 box and containing 78260 atoms, 23885 water molecules, 67 Na^+ and 75 Cl^- ions. The EGF system was solvated in a $67.0 \times 70.8 \times 117.0$ Å^3 box and containing 52417 atoms, 16208 water molecules, 47 Na^+ and 46 Cl^- ions. System relaxation was carried out using the protocol described in ref.[16]. Energy minimization and thermalization were conducted under NPT conditions at 1 atm and 298 K using a time step of 2 fs for energy minimization and a time step of 4 fs for thermalization, a cutoff of 9 Å, with rigid bonds and PME for long-range electrostatics with grids of $80 \times 76 \times 138$ for the cetuximab system and $68 \times 72 \times 118$ for the EGF system. Potential energy minimization was run for 2 ps to and thermalization for volume relaxation was run for 1 ns. During minimization the heavy protein atoms were restrained by a $1 \mathrm{kcal\,mol^{-1}\,Å^{-2}}$ spring constant and during thermalization only Cα atoms were restrained. Preparation simulations were run using ACEMD[22] on local GPU-equipped workstations.

TABLE I: Binding free energies (ΔG°) from experimental measurements (exp) and computational calculations (comp) for the EGF and FabC225 systems to the wt and S468R structures of EGFR domain III. All units are in kcal/mol.

	wt (exp)	*wt* (comp)	S468R (comp)
EGF	-7.7 ± 0.1	-6.8 ± 0.5	-7.9 ± 0.6
FabC225	-11.9 ± 0.1	-9.8 ± 0.3	-8.8 ± 0.4

S468R mutants of EGFR were generated on the basis of the wt EGFR crystallographic model. Although original mutation is reported as S492R, we have kept residue numbering as in the crystal structure solved by Li et al.[6]. Mutant system for FabC225 was simulated for 1 μs each at a temperature of 310 K in NVT conditions. The antibody conformations closest (lowest rms deviation) to the average sampled conformations were taken as starting structures for binding affinity calculations.

Binding affinity calculations Production simulations were run using ACEMD on GPUGRID.net[23] with the same parameters used for the thermalization but a time step of 4 fs using the hydrogen mass repartition scheme[22,24]. This scheme allows for longer time steps mathematically preserving all the equilibrium properties of the system, while providing only minor changes in the transport properties. Binding affinity calculations were performed using a previously reported protocol based on a one-dimensional potential of mean force reconstructed from umbrella sampling simulations[15,16]. Each umbrella sampling calculation was composed by 25 windows that ran for 50 ns for both systems. A total of four different systems were simulated: FabC225–wtEGFR, FabC225–S468R EGFR, EGF–wtEGFR and EGF–S468R EGFR. Five different replicates were run per each window and system which made up for aggregates of about 6.25 μs of MD data for each system. Final absolute binding free energy values do not incorporate the first 30 ns of data for each window, considered equilibration time (see Figure 4).

III. RESULTS AND DISCUSSION

The reported loss of treatment efficacy by cetuximab against S468R EGFR is likely to be caused by a direct disruption of the binding affinity of the drug for the receptor. On the other hand human monoclonal antibody panitumumab does not suffer the same consequences[10]. Unfortunately however, since no crystallographic structure is available for panitumumab we are unable to pro-

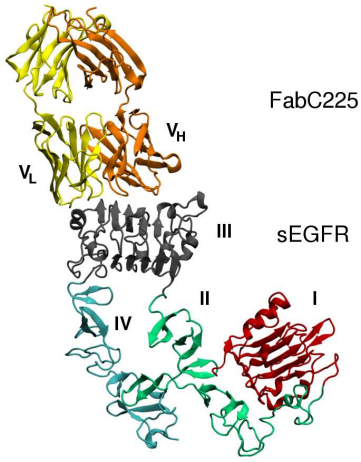

FIG. 2: Complex structure for the bound FabC225 (cetuximab) and single chain EGFR as crystallized by Li et al.[6]. Computational modeling was performed on the binding domains of both partners, VL and VH for FabC225 and domain III for EGFR since it is the only domain affected by the described mutation.

TABLE II: Estimated (est) changes in binding affinities (K_D) for EGF and FabC225 to the full S468R EGFR receptor based on the calculated binding free energies of Table I. Fold-changes are computed from reference experimental affinities[6] for EGF and FabC225 to full wt EGFR, which are 130 nM and 1.7 nM, respectively. All units are in nM.

	fold-change	S468R (est)
EGF	×2.7 increase	48.1
FabC225	×5.0 decrease	11.5

vide a structural explanation of the differentiated response between the two antibodies. Nevertheless, we have performed a structural study of the site of the mutation for the endogenous EGF-like ligands of EGFR, as well as for cetuximab and an alternative anti-EGFR human monoclonal antibody named necitumumab. Moreover, we computed binding free energies of cetuximab and EGF for the wild type and mutant EGFR structures to quantify the effect of the S468R mutation.

The crystallographic structure of FabC225 (cetuximab) in complex with the soluble extracellular sEGFR shown in Figure 2 reveals a single interaction interface between the drug and the target. The interface is vastly that of the endogenous EGFR ligands which makes cetuximab a competitive inhibitor to receptor-activating ligands[6]. Modification of these interfaces has the potential to affect complex formation and, as a matter of fact, as shown in Figure 3, the new S468R mutation may have a different impact for the binding of the drugs or the ligands. The site of the S468R mutation lies right in the middle of the surface recognized by cetuximab as shown in Figure 3a and very close to the C-terminal of EGF and TGFa and, presumably, right underneath of a putatively bound HRGa (Figure 3b-d). In the case of EGF, an additional salt bridge may become possible between E51 and S468R. In HRGa the number of additional possible interactions doubles, between E57, E61 and S468R. For TGFa as well as other EGF-like ligands like Epiregulin, not shown, the mutation is not expected to have any effect on the binding affinity of the ligand.

Since no crystal structure is currently available for panitumumab, we visually compared the cetuximab interface with EGFR with necitumumab, an alternative anti-EGFR antibody that has a very similar epitope to cetuximab[11]. Figure 5 shows the structures of FabC225 (cetuximab) and Fab11F8 (necitumumab) with respect to the mutation site S468. We have visually assessed the impact of the S468R mutation on both structures. The complexity of the interaction interfaces is such that mutations might have very different consequences on the affinity of the complexes, as indirectly seen for panitumumab[10]. The missense S468R mutation is an amino acid substitution, Serine to Arginine. Such mutation involves a change from a rather small, polar and uncharged side chain in Serine to a large and electrically charged side chain in Arginine. Two drastic changes that combined, may have deleterious effects in maintaining tight hydrophobic interactions and shape-complementary in protein-protein interfaces. Electrostatic potential calculations on the surface of EGFR domain III showed a dominating presence of positive charge[11]. The substitution of a Serine by an Arginine should favor the positively charged environment by establishing salt bridges or hydrogen bonds with the antibody although it doesn't seem to be the case. Moreover, the two antibodies, cetuximab and necitumumab might be differently affected by the mutation. As shown in Figure 5b the principal differences between FabC225 and Fab11F8 are the presence of residues Y104, W52 and W94 in FabC225 bound near the S468 in EGFR. Residue Y104 in particular, appears to be obstructing an otherwise accessible cavity for Arginine. Overall, the addition of a large and charged amino acid may cause a costly side chain rearrangement of cetuximab residues near S468 together with increased solvation that would would impede tight complex formation characteristic of antigen-antibody interfaces.

In order to determine a putative decrease of binding affinity for the FabC225–EGFR complex and a putative increase for the EGF–EGFR complex, we performed computational binding free energy calculations[15,16] of the two complexes using high-throughput all-atom molecular dynamics simulations[23]. Table I shows a summary of the calculated binding affinities for FabC225/EGFR domain III and EGF/EGFR domain III both for their wild type and mutant forms. EGF was found to bind to EGFR domain III with a free energy of -6.8 ± 0.5 kcal/mol (compared to an experimental of -7.7 kcal/mol[11]) and FabC225 with -9.8 ± 0.3 kcal/mol (compared to an exper-

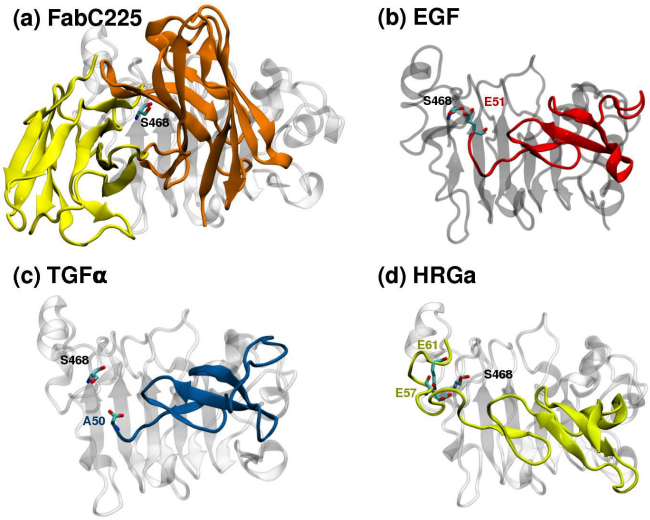

FIG. 3: Spatial relationship between the mutation site S468 in EGFR and the bound structures of FabC225 and several receptor-activating ligands. (a) S468R mutation may have a clear impact of the binding of FabC225. A complex number of surrounding interactions may be affected with the inclusion of a large and charged amino acid as it is shown in Figure 5. (b) In the case of EGF, the mutation may on the other hand, increase the affinity of the ligand. A new salt bridge interaction may exist upon mutation of Serine to Arginine with E51 in EGF. (c) TGFa has A50 in close proximity to mutant site. Binding affinity might be slightly increased after mutation S468R by interaction between the C-terminal of TGFa and Arginines side chain. (d) HRGa is also displayed for reference. A non crystallographic binding mode has been modeled to asses an hypothetical interaction between S468R and E57/E61 in the ligand. Although truncated in this figure, non-native ligand EGFR ligand HRGa may gain in binding affinity for S468R EGFR being an additional competitor to a weaker binder cetuximab.

imental of -11.9 kcal/mol[6]). Considering the size of the system, the accuracy of these calculations for the *wt* complexes is remarkably high, specially for EGF which is less than 1 kcal/mol off from the experimental value, being this difference perfectly equivalent to the ones reported in previous work for a tetrapeptide ligand on the same protocol[16]. As expected for the S468R mutant complexes, calculations predict a binding free energy 1.1 kcal/mol more favorable for EGF and 1 kcal/mol less favorable for FabC225, although the latter is a less reliable result given the oversimplification of the simulated model that used only the Fv part of the antibody. Binding free energies reported in Table I for FabC225 are the mean and standard deviation of the 5 replicas per system where each replica value is obtained from the latest quarter of sampled time and from the last 10-20 ns sampled in EGF (see Figure 4 for convergence studies). The calculated free energy values however, only considered interactions with domain III of EGFR but EGF, for instance, is known to bind with greater affinity to full EGFR since it also interacts with domain I[17]. Assuming that domain I contributes equally to the total measured binding affinity, in Table II we show the final estimated binding affinities for the full mutant S468R EGFR taking into account the free energy calculations. Cetuximab is predicted to display at least a 5-fold decrease in binding affinity for S468R EGFR and EGF is estimated to display a binding affinity increase of 2.7-fold.

IV. CONCLUSION

None of the reported mutations *in vitro*[6,11] can individually match the deleterious effect that acquired mutation S468R displays in FabC225 binding. Although highly significant for a single residue mutation, a 5-fold decrease in binding affinity may not be enough to cause the described resistance seen in the treatment[10]. It may be the combination with the 2.7-fold increase in EGF binding affinity and increased putative competition by other EGF-like ligands like HRGa that impedes receptor inhibition *in vivo*. Most of the mutations that have been explored in EGF binding to EGFR domain III actually caused a decrease in affinity[11]. Only the combined mutation Q408A/H409A has a significant increase of 2.7-fold in EGF binding affinity[11]. In this work, we show how a single missense mutation can cause both a decrease in drug binding and an increase in endogenous ligand bind-

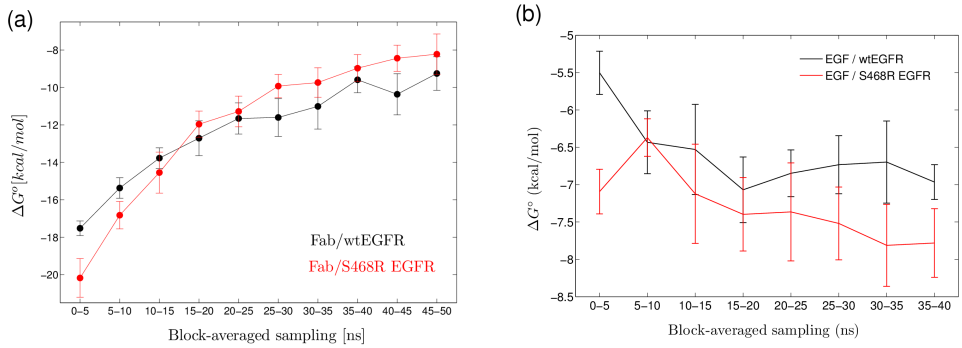

FIG. 4: Binding free energy (ΔG°) convergence studies for the (a) cetuximab system and (b) EGF system versus single umbrella sampling window simulated time. Free energy values are computed as the mean and standard deviation of the 5 different replicas across block-averaged time ranges to assess convergence. In subplot (a) although free energy values seem to reach a plateau phase at 50 ns, we cannot confidently discuss their convergence. Differences between the wt and mutant systems are of about 1 kcal/mol only. Nevertheless, it is very likely that the modeled system is an inaccurate oversimplification due to an excessive reduction of the antibody domains simulated since only the Fv part was modeled. In subplot (b) free energy values for EGF against wt and mutant EGFR are clearly differentiated specially after 30 ns of sampling time per window. In the computation of the absolute binding free energies reported, the first 30 ns of each window were discarded for the computation.

ing. Both changes are predicted to render binding affinities of the same order of magnitude for the two complexes (48.1 nM for EGF and 11.5 nM for cetuximab) which may be the reason of the failure of the therapeutic strategy versus S468R mutant EGFR.

Although the accuracy of the FabC225 model is certainly problematic due to the simplified modeling strategy, the differential binding free energy values obtained for the wild type and the mutant receptors may still be a fair approximation to the actual affinities of the full antibody chains. On the other hand, we are very confident on the the results for the EGF system which showed a remarkable precision and accuracy given the size and complexity of the system. Ultimately, these results need to be validated with experimental measurements of binding affinities and kinetics for the mutant EGFR

Finally, this work is an example of how in the near future and in the context of personalized medicine, binding free energy calculations could be successfully used to predict the efficacy of existing drugs to unknown target variants.

[1] WHO (2012) Who cancer factsheet. (http://tinyurl.com/6fg6ofz).

[2] McKay JA, et al. (2002) Evaluation of the epidermal growth factor receptor (EGFR) in colorectal tumours and lymph node metastases. *European Journal of Cancer (Oxford, England: 1990)* 38:2258–2264 PMID: 12441262.

[3] Porebska I, Harlozińska A, Bojarowski T (2000) Expression of the tyrosine kinase activity growth factor receptors (EGFR, ERB b2, ERB b3) in colorectal adenocarcinomas and adenomas. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine* 21:105–115 PMID: 10686540.

[4] Hoy SM, Wagstaff AJ (2006) Panitumumab: in the treatment of metastatic colorectal cancer. *Drugs* 66:2005–2014; discussion 2015–2016 PMID: 17100412.

[5] Yang XD, Jia XC, Corvalan JR, Wang P, Davis CG (2001) Development of ABX-EGF, a fully human anti-EGF receptor monoclonal antibody, for cancer therapy. *Critical Reviews in Oncology/Hematology* 38:17–23 PMID: 11255078.

[6] Li S, et al. (2005) Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell* 7:301–311 PMID: 15837620.

[7] Prewett M, et al. (1996) The biologic effects of c225, a chimeric monoclonal antibody to the EGFR, on human prostate carcinoma. *Journal of Immunotherapy with Emphasis on Tumor Immunology: Official Journal of the Society for Biological Therapy* 19:419–427 PMID: 9041461.

[8] Klastersky J (2006) Adverse effects of the humanized antibodies used as cancer therapeutics. *Current Opinion in Oncology* 18:316–320 PMID: 16721123.

[9] Bardelli A, Jänne PA (2012) The road to resistance: EGFR mutation and cetuximab. *Nature Medicine* 18:199–200.

[10] Montagut C, et al. (2012) Identification of a mutation in the extracellular domain of the epidermal growth factor receptor conferring cetuximab resistance in colorectal cancer. *Nature Medicine* 18:221–223.

[11] Li S, Kussie P, Ferguson KM (2008) Structural basis for EGF receptor inhibition by the therapeutic antibody IMC-

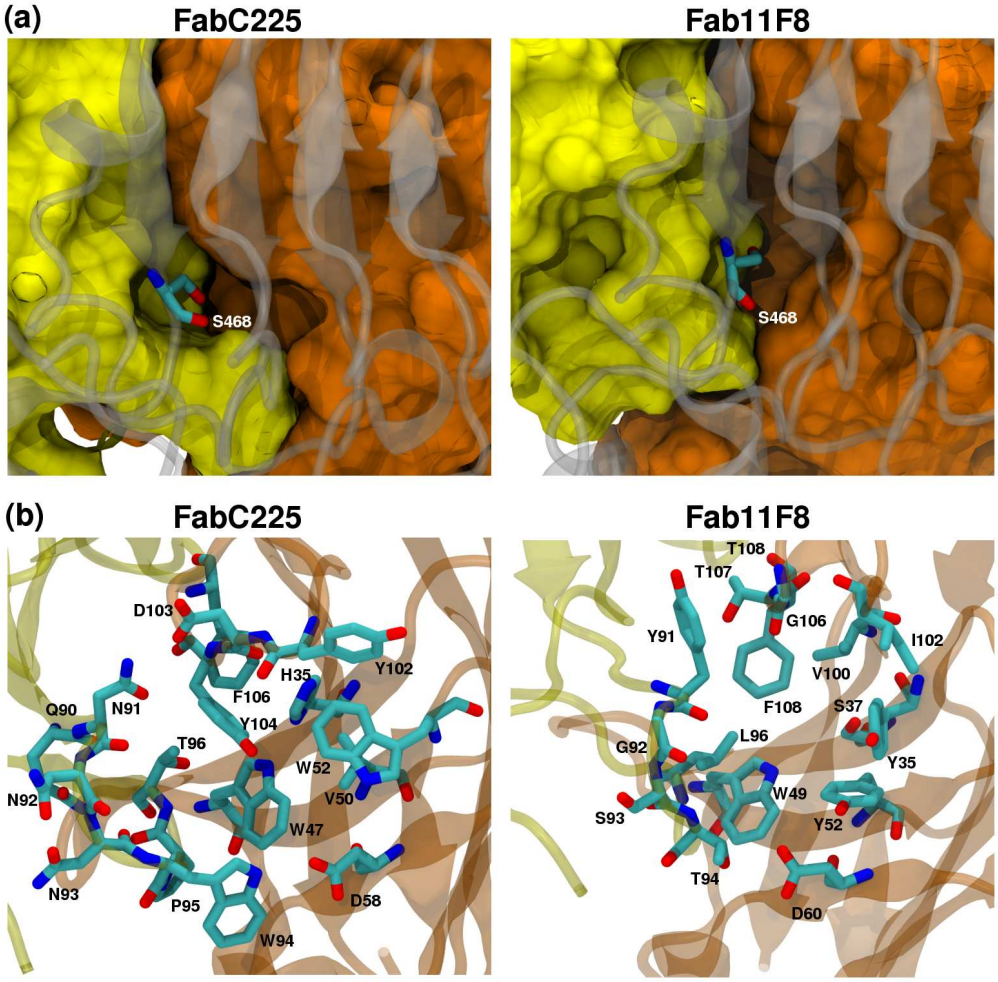

FIG. 5: Assessing a potential accommodation of mutation S468R in mAb interacting interfaces. (a) Compared to FabC225 (cetuximab), Fab11F8 (necitumumab) displays a cavity that may be able to accommodate a bulky residue such as Arginine. (b) A selection of residues between 5-7 Åof S468 in the crystal structures show a the residues responsible for the different antibody surfaces. Principal differences between FabC225 and Fab11F8 are residues Y104, W52 and W94 in FabC225. Indeed, residue Y104 in FabC225 practically obstructs an otherwise accessible cavity for Arginine. Tryptophan side chains parallel to the binding interface contribute to a strong hydrophobic interaction. The cost in a major side chain rearrangement in cetuximab residues near S468 may be incompatible with the conservation of a tightly bound antibody. Moreover, the addition of a charged amino acid and a slight separation of the complex, may facilitate the entrance of water molecules and difficulting tight complex formation characteristic of antigen-antibody interfaces. The more accessible Fab11F8 cavity near S468 may be more likely to accommodate a mutant Arginine and therefore display a less negative response to the mutation.

11F8. *Structure* 16:216–227.

[12] Woo H, Roux B (2005) Calculation of absolute protein–ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* 102:6825–6830.

[13] Gervasio FL, Laio A, Parrinello M (2005) Flexible docking in solution using metadynamics. *Journal of the American Chemical Society* 127:2600–2607 PMID: 15725015.

[14] Jorgensen WL (2004) The many roles of computation in drug discovery. *Science* 303:1813 –1818.

[15] Doudou S, Burton NA, Henchman RH (2009) Standard free energy of binding from a One-Dimensional potential of mean force. *Journal of Chemical Theory and Computation* 5:909–918.

[16] Buch I, Sadiq SK, De Fabritiis G (2011) Optimized potential of mean force calculations for standard binding free energies. *Journal of Chemical Theory and Computation* 7:1765–1772.

[17] Ogiso H, et al. (2002) Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* 110:775–787 PMID: 12297050.

[18] Garrett TP, et al. (2002) Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor . *Cell* 110:763–773.

[19] Jacobsen NE, et al. (1996) High-Resolution solution structure of the EGF-like domain of heregulin-. *Biochemistry* 35:3402–3417.

[20] Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Research* 28:235–242 PMID: 10592235.

[21] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79:926–935.

[22] Harvey MJ, Giupponi G, Fabritiis GD (2009) ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *Journal of Chemical Theory and Computation* 5:1632–1639.

[23] Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G (2010) High-Throughput All-Atom molecular dynamics simulations using distributed computing. *Journal of Chemical Information and Modeling* 50:397–403.

[24] Feenstra KA, Hess B, Berendsen HJC (1999) Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of Computational Chemistry* 20:786–798.

3.4 Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations

Buch I., Giorgino T. and De Fabritiis G., *Proceedings of the National Academy of Sciences of the USA* 108, 10184–10189 (2011)

Summary

In this work we exploited the power of GPUGRID to quantitatively reconstruct the complete unbiased binding process of the enzyme-inhibitor complex trypsin-benzamidine. Through the simulation of 495 molecular dynamics trajectories of free ligand binding of 100 ns each, we obtained 187 binding events with an RMSD less than 2 Å compared to the crystal structure that allowed us to reconstruct the binding pathway and estimate the binding free energy and rates. We have identified previously unknown metastable intermediate states for the binding of benzamidine to trypsin that highlight potential key residues in the kinetics of benzamidine binding. The estimation of the standard free energy of binding gives $\Delta G^{\circ} = -5.2 \pm 0.4$ kcal/mol (cf. the experimental value -6.2 kcal/mol), and the binding kinetic rates $k_{on} = (1.5 \pm 0.2) \times 10^{8}\,\mathrm{M}^{-1}\mathrm{s}^{-1}$ and $k_{off} = (9.5 \pm 3.3) \times 10^{4}\mathrm{s}^{-1}$ for unbound to bound transitions. With this hallmark piece of work we demonstrate the predictive power of unconventional high-throughput molecular simulations, as well as introduce a methodology that is directly applicable to other molecular systems and thus of general interest in biomedical and pharmaceutical research.

Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci USA. 2011 Jun 21;108(25):10184-10189.

Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Supporting information. Proc Natl Acad Sci USA. 2011 Jun 21;108(25):10184-10189.

3.5 Visualizing the induced binding of SH2-phosphopeptide

Giorgino T., Buch I. and De Fabritiis G., *Journal of Chemical Theory and Computation*, 8, 1171–1175 (2012)

Summary

Following on with the unbiased ligand binding simulations approach, here we report in atomistic detail the way a phosphorylated peptide binds to the ubiquitous SH2 domain and the conformational changes that take place upon binding. To do so, we obtained several spontaneous binding events between the p56 lck SH2 domain and the pYEEI peptide within 2 Å RMSD from the crystal structure and with kinetic rates compatible with experiments using high-throughput molecular dynamics simulations. We describe how binding is achieved in two phases through first, fast contacts of the charged phosphotyrosine and second, then rearrangement of the ligand involving the stabilization of two important loops in the SH2 domain. These observations provide insights into the binding pathways and induced conformations of the SH2–phosphopeptide complex which, due to the characteristics of SH2 domains, should be relevant for other SH2 recognition peptides. On a broader perspective and provided that sufficient sampling was provided, this work is ultimately relevant as an aid to the reconstruction of complex recognition models via Markov state models.

Giorgino T., Buch I. and De Fabritiis G. Visualizing the induced binding of SH2-phosphopeptide. J Chem Theory Comput. 2012; 8: 1171 -1175.

Giorgino T., Buch I. and De Fabritiis G. Visualizing the induced binding of SH2-phosphopeptide. Supporting information. J Chem Theory Comput. 2012; 8: 1171 -1175.

Chapter 4

DISCUSSION

The following discussion deals with the overall impact of the results obtained and their contextualization within the current state of the art.

Setup and development of GPUGRID for high-throughput molecular dynamics simulations

GPUGRID started in 2007 as a volunteer distributed computing grid of PlayStation3 named PS3GRID thanks to the then-innovative "Cell" multiprocessor [200]. Transition to GPUs in 2008–09 made the project become GPUGRID, receiving the total contribution of more than 15000 users that volunteered together more than 28000 computers. The active percentage of these users as of March 2012 is around 17%. This makes for an active number of contributed GPUs of about 3000. The mixture of a high-performance architecture like the GPUs and the high number of contributing users, made GPUGRID one of the top players in the distributed computing community for biomedicine, sharing the ranking with projects from renowned institutions and research programs such as Stanford University's Folding@Home [82] (400,000 volunteers) or Washington State University's Rosetta@Home [201] (36,000 active volunteers) too.

We can identify two main elements in the development of GPUGRID that have been key in making it both a successful tool for researchers and an attractive distributed computing project for volunteers. For GPUGRID to be a practical tool for the every day usage by the scientist, it needed to fix interfacing issues with the rather complex BOINC server software with regard to simulation submission and data retrieval. To solve this, Giorgino et al [169] developed the

RBoinc interface that has represented a critical advancement for the seamless utilization of GPUGRID as a supercomputing platform and in particular, the daily submission and retrieval of computing tasks, named work units.

The other aspect that has made of GPUGRID an attractive volunteer computing project has been the high standards achieved in managing the community of volunteers. In particular, the efforts put in the daily follow-up of the volunteer's concerns on the project forums, the constant development of a usable and attractive project website (currently in its third version) as a source for scientific information on the goals of the project and, finally, the implementation of a unique and distinctive visual contribution recognition system, a 'badge system'. The objective of the latter is granting visible recognition to users according to their proportional contribution. This is of great importance since improvements in the website layout and in public visibility typically have an immediate and tangible impact on the number and contribution of volunteers in distributed computing projects (Figure 1.8) [170, 202].

An important aspect of volunteer distributed computing is the democratization of science. As the father and current coordinator of BOINC put it, when computer owners can contribute to whatever project they choose, control over resource allocation for science may be shifted away from government funding agencies and towards the public. Such shift in control comes with the risk of volatility of public interest but in turn it offers a very direct and democratic mechanism for deciding research policies [203]. A different question is then whether the scientific community is ready yet to embrace such levels of democratization. Nevertheless and as far as we are concerned, no distributed computing project has been yet forced by its volunteer-base to a shift in their research focus. From the scientists point of view, maintaining a fluent and open attitude towards the communication of the scientific projects' nature is usually sufficient to content the interested nature of the majority of the volunteers.

As a scientific computing infrastructure, GPUGRID enabled us to routinely sample microseconds of data, while multi-μs experiments were previously a rarity. This is specially relevant considering that biological phenomena, such as ligand binding or large conformational motions, start to occur at the microsecond timescale [64]. Costly but highly parallel simulation protocols like free energy calculations from one-dimensional potential of mean force [81, 121, 71], have been successfully implemented during this time. Although its initial application was intended for the computation of binding free energies by surmounting

the ‘sampling problem’, GPUGRID is now also being employed to capture and quantify rare dynamic molecular events. Events that happen at characteristic rates and therefore, by increasing the number of parallel runs and thus of total simulated time, we can increase our chances to capture them. These two methodologies were partially motivated and developed as a refinement of the work of this thesis. Thus, GPUGRID has so far allowed simulations of binding of protein–ligand systems [135, 27], ion-induced effects in GPCR molecules [119] and rare conformational state characterization in HIV-1 protease [204].

Considering the state-of-the-art of MD computing, GPUGRID is, together with Folding@Home for protein folding [148], a leading infrastructure in the world in the application of high-throughput MD simulations. The amount of sampling that is daily produced is in close competition with DE Shaw Research, home of several Anton, the multimillion-dollar special-purpose machines [79]. Their approach however, differs from our high-throughput and is focused on extreme high-performance aimed at simulating single very long MD trajectories mainly of protein folding [67, 205] and ligand binding [206, 207].

Implementation and application of a one-dimensional potential of mean force-based method for binding free energy calculations

Attempts to compute binding affinities have been made since near the inception of computational biomolecular modeling [208]. In the Introduction section of this thesis, we have already reviewed how several methods involving MD simulations are being used to quantitatively determine binding free energies [106, 117, 124] as well as their approximations to physical representation of binding [209, 128]. The underpinning limitations of these methods are still essentially related to the computational cost of the estimations and the requirement of expert human intervention. These are precisely the two main limitations that we wanted to tackle in computing binding free energies with GPUGRID.

Focusing on all-atom physical pathway-based MD simulations, Roux and co-workers [124, 210] introduced PMF-based one-dimensional absolute binding affinity calculations. Their approach was motivated by the need to tackle more complex situations in biological systems dealing with, for example, flexible and charged ligands [124, 209]. The method, based on umbrella sampling, solved the large computational cost associated by the application of a set of conformational and oriental restraining potentials to the ligand. The study claims that a few nanoseconds of sampling are sufficient to obtain converged results

for the free energy. This approach however, although apparently solving the sampling problem, is hard to generalize to regularly compute binding free energies for other systems, because it requires a deep knowledge on the degrees of freedom of the system that play a role in the recognition. This is a general problem of biased PMF-sampling methods, where their efficacy and efficiency are determined by the choice and number of coordinates of reaction [211]. Metadynamics, for example, despite being one of the most popular pathway-based sampling methods [117, 118, 212] has traditionally suffered from the very same problem of having to decide beforehand the relevant degrees of freedom for the system, namely the collective variables. This decision, in turn, also affected the choice of simulation parameters and the convergence of the free energy estimates [213, 214] which are also common problems in other pathway-based methods [211, 121, 134].

An important step forward with respect to the methodology proposed by Roux and co-workers, was the work by Henchman and co-workers [134] that presented an updated version of the one-dimensional PMF protocol. They removed conformational and rotational restraints on the ligand and left only generic restraints, orthogonal to and in the direction of binding. Moreover, they provided the framework for the calculation of standard free energy of binding, something often overlooked in comparing calculations to experiment [117, 32]. This version of the methodology had a major improvement in the ease of implementation leaving only the question of the sufficient sampling to be resolved. Regarding the definition of the single coordinate of reaction, an obvious choice for protein–ligand binding is the distance between the two. The coordinate is often chosen to be either the radial distance or a projection to a cartesian axis orthogonal to the binding interface of the complex [124, 134].

In our implementations, there were two main differences from the original protocol that addressed the setup and convergence issues. First, the utilization of steered MD simulations to slowly displace the ligand away from the protein along the chosen reaction path orthogonal to the interface [81]. The snapshots of this 'pulling' run served as initial configurations for the umbrella sampling. In this way, only a single thermalization run for the system had to be performed and whole composition of the simulation box was maintained. Also, to some extent, we were generating trajectories across unbinding pathways that could have been relevant for binding as well. The second difference in the implementation of the protocol consisted in using uncorrelated starting configurations for neigh-

boring umbrella sampling windows and it was incorporated in the optimization presented in publication 3.2.

In the first application of the protocol we calculated a converged standard free energy of binding for the SH2–pYEEI [215] system of -8.7 ± 0.4 kcal/mol within 0.7 kcal/mol from experimental results, however at the cost of 20.5 μs of data [81] which made the protocol unfeasible to regular and generalized application. To solve this, we optimized the protocol cutting down the computational cost to only 300 ns for the same protein–ligand system to achieve an even more precise value of -9.0 ± 0.5 kcal/mol, despite its larger error (1 kcal/mol). In this optimization study, presented in publication 3.2, we were able to propose an optimized version of the one-dimensional PMF protocol based on umbrella sampling that consisted in using an ensemble of simulations, initiated from uncorrelated initial conformations across neighboring windows and an optimal parameter set (OPS) describing orthogonal restraints, a force constant for the sampling potential, window width, and sampling time per window.

In publication 3.2 we additionally provided an example case of insufficient sampling that related overestimated and underestimated binding free energies with metastable structural correlates at the transition region of the binding/unbinding pathways. This observation was in agreement with previous work by Mobley et al. [216] where they stated that conformational changes can make a difference of several kcal/mol in computed binding free energies, and that free energy estimations of systems kinetically trapped in particular metastable states can incur in large estimation errors. Others have also arrived to similar conclusions in the computation of absolute binding free energies on various systems with significant degrees of flexibility [217, 218, 219]. Far from being universally optimal, our protocol is certainly best performing when feeding the umbrella sampling simulations with numerous and uncorrelated variety of starting configurations. Ultimately, the work on SH2 has been an important test case given the degree of flexibility of both the ligand and the protein as we also saw in posterior work on unbiased binding simulations for the same system [27].

In a more ambitious study and along these lines, we have lately applied the protocol to a protein–ligand system of much larger size albeit similar flexibility, the EGFR–cetuximab and EGFR–EGF systems. In publication 3.3 we try to provide a binding affinity-based explanation for the resistance to the monoclonal antibody drug cetuximab, to the recently described mutant variant S468R of EGFR [199]. In addition to making a structure-based assessment of the pu-

tative impact of the mutation to the stability of target–drug and target–ligand complexes, we performed extensive binding free energy calculations to specifically determine the impact of the mutation in the binding of cetuximab and EGF. Our calculations suggest that, unlike for other previously described mutations for the system [52, 220], there would be a strong deleterious effect of the treatment efficacy from a simultaneous 5-fold decrease of binding affinity for cetuximab and a 2.7-fold increase in EGF binding affinity which may finally be impairing competitive binding from the drug. Similar effects, upon mutation of a single residue, have also been described for single residue mutations in the intracellular kinase domain of EGFR [221]. Although our predictions still need to be experimentally validated, as well as to determine the specific effects on the kinetic rates for the drug and other EGFR ligands, the overall conclusion is in line with the phenomena reported from the clinical and *in vitro* studies [199].

Given the computational cost associated, the one-dimensional PMF-based protocol would still not be practical in screening stages of drug discovery. Instead, it may have a role to play in later-stages at lead optimization or for screenings of emerged resistances to approved drugs in the direction of personalized therapies [222, 223]. Moreover, it might be best suited to the study of biologicals like peptides or antibodies where conformational flexibility is more problematic.

Implementation and application of unbiased sampling methods for complete binding process reconstruction

Without doubt, the grand challenge in the study of protein–ligand interactions is the direct observation and quantification of unbiased equilibrium-based ligand binding at atomic resolution, something which has remained at a prohibitive computational cost until now. Although some spontaneous binding events had already been reported [224, 225, 75], the first statistically meaningful binding experiments appeared concomitantly with our work on trypsin-benzamidine binding of publication 3.4. Shaw and co-workers presented the longest-ever simulated binding trajectories using all-atom MD simulations of kinase inhibitors dasatinib and PP1 binding to Src kinase [207] and several other inhibitors binding to β1- and β2-adrenergic GPCR receptors [206]; in both works they provide estimations for the association rates and some binding free energy but, unlike us, they were unable to report estimates for ligand dissociation rates. On the other hand, Silva et al. [152] providing full description of the binding process with kinetics, affinity and pathway in LAO protein binding and amply discussed

the roles for conformational selection and induced-fit. More recently, in publication 3.5, we presented an extensive unbiased phosphopeptide binding study to SH2 that produced 5 binding events out of 772 trajectories due to a slow and complex kinetic mechanism. Without being able to discuss the roles for conformational selection or induced fit due to insufficient sampling, we were able to provide a dynamic view of the conformational flexibility of SH2 and its relationship to phosphopeptide binding. In summary, all these ground-breaking works are mainly a consequence of maturity of high-performing codes and architectures [79, 162, 80] and, in some cases, advances in the development of ensemble-based transition network analysis methods like Markov state modeling (MSM) [144, 146].

Unbiased binding simulations have the advantage of not having to assume the coordinate of reaction in advance which otherwise may provide a biased or oversimplified view on the kinetics [154], instead, study of the relevant degrees of freedom is performed afterward [145]. This fact added to the ability to reconstruct the equilibrium ensemble of binding pathways with MSM from simulations that are much shorter than the binding time [147], makes of high-throughput unbiased binding MD simulations an ideal next-generation approach to investigate protein–ligand binding. Indeed, MSM is a suitable mathematical framework to analyze GPUGRID-produced unbiased sampling data. Before MSMS we had been studying biological systems with an approach based on the simple premise of 'simulate (unbiasedly) and see what happened'. Whether it has been ligand binding [71, 27] or protein conformational changes [204, 119] we have been using human intuition and manual projections to capture and hopefully quantify conformational dynamics of rare events. Where others might use PCA analysis [226] or complex algorithms like sketch-map [227] to extract the collective degrees of freedom for dimension-reduced descriptions of macromolecular dynamics, we are now directly using MSM in an iterative manner to find those dimensionality reductions that better describe the dynamics of our system.

The theory of MSM is sound and well developed [228, 144, 145, 146]. In particular, the formalisms that allow the computations of the the statistical quantities of the ensemble which have a direct meaning in MD: the equilibrium distribution of the system, kinetically meaningful metastable states and transition rates between these states [146]. For binding affinity calculations, for example, we can alternatively compute standard free energy of bindings through integration of the PMF [134] as we had been doing from umbrella sampling simulations or through

the association and dissociation rates of the system from which the equilibrium constant can be calculated, as expressed in equation (1.2). Also, in a more complex implementation of the MSM analysis one could build complex kinetic network and represent the transitions between the different metastable states from the calculation of their relative populations in equilibrium, the free energies, to their rates and weights of interconversion, the fluxes [148, 229, 147, 150]. Such a deep analysis, although unnecessary in the simple quasi-binary process of trypsin-benzamidine [135], should certainly be undertaken in further developments of the more complex binding mechanisms that we described for the SH2–pYEEI binding [27] in a similar way to work by Silva et al [152].

More specifically, in our first application of MSMs presented in publication 3.4, we obtained 187 full binding events out of 495 analyzed trajectories. Such a large number of sampled transitions permitted rather precise estimations for binding affinity and association rate, but not so much for dissociation rates. Although we did not obtain a single full unbinding event, the MSM was able to predict the unbinding rate but overestimated by two orders of magnitude. Additionally, we provided information of the binding pathway highlighting metastable binding sites as well as transitory interactions on the surface of trypsin, that participated in process of benzamidine binding to the canonical pocket. Recently, some of these findings have been reproduced using alternative biased sampling methodologies like reconnaissance metadynamics for binding pose discovery [230]. Parrinello and co-workers identified some of the metastable states and transient interactions we had described for benzamidine on the surface of trypsin, named S2, S3 and TS1-TS3. We are also currently extending the methodology to the discovery of alternative binding sites in collaboration with researchers at the European Synchrotron Radiation Facility in Grenoble who are able to obtain crystallographic structures of short-lived protein-ligand complexes using cryoprotectant-free high-pressure freezing [231]. Indeed, being able to find ligand binding poses, either canonical or alternative, has the promising potential of aiding the design of allosteric modulators targeting these sites; already a declared driving motivation for some of the recently published studies [207, 206]. Moreover, these alternative of metastable states also have the potential to provide information on the kinetic properties of target–drug interactions [43].

As already mentioned in the Introduction section, kinetics of binding is gaining attention in drug discovery as it has been described to generally provide better correlations with *in vivo* drug activities than binding affinities [38, 232].

The analysis from Swinney [40, 41] revealed that an increasing amount of drugs approved by the FDA had non-equilibrium kinetics and induced conformational changes in proteins. Moreover he suggested that rapid dissociation rates are a means of minimizing mechanism-based side effects [42]. In general, consideration of association (k_{on}) and dissociation (k_{off}) rates of binding on the design of drugs may thus have important contributions to the efficacy, safety, duration of action and differentiation of these drugs [40]. In recent work, Barril and coworkers [43] have recently demonstrated that formation of water-shielded hydrogen bonds between a ligand and its receptor protein increases the kinetic stability of complexes. Control over kinetic structure activity relationships is set to be one of the next major goals in drug discovery.

If the promising role of kinetics in drug activity is confirmed, the combined approach of high-throughput MD simulations with MSM analysis could soon become a revolutionary tool in the context of structure-based drug discovery. Although the capabilities are still very much limited to fast associating ligands, it could soon be made more generally applicable with the development of adaptive MSM strategies; an improvement over the standard MSM that allows for adaptively enhancing sampling in insufficiently-resolved transitions [146, 233, 148]. Moreover, there is still ample space for learning and controlling the effects that several parameters in the building of MSM have on the convergence and the error on statistical quantities that the method is able to provide [146].

Finally, high-throughput MD simulations with MSM analysis may have a future application in structure-based drug discovery on the specific sub-discipline of fragment-based design and discovery. The principal idea behind fragment-based drug discovery is to increase the probability of finding hits in libraries of small-sized ligands or even molecular features [234]. In a sort of Lego-like approach, through the covalent combination of neighboring fragment hits on a target, commonly known as 'growing fragments', highly potent and selective drugs can be designed. Drugs that would have otherwise not been present in common libraries [235, 236]. In this context, high-throughput MD fragment binding simulations may be the means by which hits or leads could be screened and ranked *in silico* for affinity and kinetics becoming an all-in-one solution for fragment-based drug discovery.

Chapter 5

CONCLUSIONS

1. Volunteer distributed computing is a cost-effective alternative for high-throughput scientific computing as long as community management stays at a comparable cost to applying for access to supercomputing facilities. Moreover, involving the society in the daily process of scientific research is an act of responsibility and may be able to democratize scientific practices and goals.

2. High-performance computing architectures like GPUs and codes like ACEMD, allow for a shorter time-to-answer and, its combination with high-throughput approaches on embarrassingly parallel infrastructures has proved capable of overcoming the sampling issue in absolute binding free energy calculations and unguided ligand binding simulations, problems long regarded prohibitive.

3. A one-dimensional potential of mean force-based binding free energy protocol to compute protein–ligand binding free energies, although still requiring extensive amounts of sampling, largely solves the need of expert human knowledge to set up calculations. Too costly for calculations of many ligands, it has been successfully applied to studies of semi-flexible protein–ligand and protein–protein complexes.

4. Molecular structure-based analysis coupled to binding free energy calculations in the determination of the impact of S468R mutation in EGFR in colorectal cancer therapy predicts that, resistance to cetuximab can be due

to both a loss in cetuximab binding affinity and a gain in EGF affinity for the receptor. Structural analysis also suggests that alternative monoclonal antibody necitumumab might be less affected by the mutation.

5. Free ligand binding allows for an unguided exploration of the conformational phase-space of protein–ligand interactions and has the potential of finding metastable binding poses that that are indicatives of putative allosteric target sites. Trypsin–benzamidine binding experiments have revealed non-obvious roles for metastable states involved in the binding pathway, away from the known native pocket.

6. Markov state modeling is able to reconstruct kinetic networks from many short unbiased simulation trajectories and quantify events with timescales several order of magnitude longer than the individual trajectories simulated. Applied to free ligand binding, MSM can readily provide a complete quantitative picture of a binding process giving binding affinity, binding kinetics and binding pathway as shown for the trypsin–benzamidine study.

7. Provided that optimal adaptive unguided sampling strategies can be successfully implemented, high-throughput free ligand binding molecular dynamics simulations analyzed with Markov state modeling may be able to play a role in future *in silico* fragment-based drug discovery enterprises. The methods and protocols implemented in GPUGRID for this purpose, can be easily ported to dedicated in-house GPU computing facilities.

Chapter 6

LIST OF COMMUNICATIONS

Articles

1. Buch I., Harvey M.J., Giorgino T., Anderson D.P. and De Fabritiis G., High-throughput all-atom molecular dynamics simulations using distributed computing, *J Chem Inf Mod* 50, 397 (2010)

2. Buch I., Sadiq S.K. and De Fabritiis G., Optimized potential of mean force calculations of standard binding free energy, *J Chem Theory Comput* 7, 1765–1772 (2011)

3. Buch I., Giorgino T. and De Fabritiis G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations, *Proc Nat Acad Sci USA* 108, 10184–10189 (2011)

4. Giorgino T., Buch I. and De Fabritiis G., Visualizing the induced binding of SH2-phosphopeptide, *J Chem Theory Comput*, 8, 1171–1175 (2012)

Oral communications

1. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations, VIII European Workshop in Drug Design, Siena (Italy), May 2011. — Awarded poster pitch

2. Energetics, kinetics and binding pathway reconstruction for enzyme-inhibitor complex from high-throughput molecular dynamics simulations,

UK Young Modellers Forum 2010, London (UK), December 2010. — Awarded talk

3. Energetics, kinetics and binding pathway reconstruction for enzyme-inhibitor complex from high-throughput molecular dynamics simulations, IV Meeting on High Performance Computing in Molecular Simulations, Madrid (Spain), October 2010.

4. Reliable and accurate prediction of ligand binding by high-throughput molecular dynamics simulations, XVIII Jornades de Biologia Molecular, Barcelona (Spain), June 2010. — Awarded talk

5. High-throughput all-atom molecular dynamics simulations using distributed computing, 24th Molecular Modeling Workshop, Erlangen (Germany), March 2010. — Awarded talk

Poster communications

1. Quantitative prediction of molecular interactions by MD simulations, I GRIB Expo, Barcelona (Spain), April 2012

2. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations, VIII European Workshop in Drug Design, Siena (Italy), May 2011. — Awarded poster pitch

3. A distributed computing system for high-throughput calculations of free energies of binding using molecular dynamics simulations, Expanding the frontiers of molecular dynamics simulations in biology, Barcelona (Spain), November 2009.

Chapter 7

BIBLIOGRAPHY

[1] McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977 Jun;267(5612):585–590.

[2] Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, et al. Global Mapping of the Yeast Genetic Interaction Network. Science. 2004;303(5659):808–813.

[3] Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996;93(1):13–20.

[4] Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003;22(14):3486–3492.

[5] Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. Structure. 2010;18(10):1233–1243.

[6] Rühlmann A, Schramm HJ, Kukla D, Huber R. Pancreatic trypsin inhibitor (Kunitz). II. Complexes with proteinases. Cold Spring Harb Symp Quant Biol. 1972;36:148–150.

[7] Amit AG, Mariuzza RA, Phillips SE, Poljak RJ. Three-dimensional structure of an antigen-antibody complex at 2.8 A resolution. Science. 1986;233(4765):747–753.

[8] Cramer P, Bushnell DA, Kornberg RD. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science. 2001;292(5523):1863–1876.

[9] Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science. 2000;289(5481):905–920.

[10] Wimberly BT, Brodersen DE, Clemons J W M, Morgan-Warren RJ, Carter AP, Vonrhein C, et al. Structure of the 30S ribosomal subunit. Nature. 2000;407(6802):327–339.

[11] Jeruzalmi D, O'Donnell M, Kuriyan J. Crystal structure of the processivity clamp loader gamma (gamma) complex of E. coli DNA polymerase III. Cell. 2001;106(4):429–441.

[12] Robinson RC, Turbedsky K, Kaiser DA, Marchand JB, Higgs HN, Choe S, et al. Crystal structure of Arp2/3 complex. Science. 2001;294(5547):1679–1684.

[13] Whittle PJ, Blundell TL. Protein structure–based drug design. Annu Rev Biophys Biomol Struct. 1994;23:349–375.

[14] Blundell TL. Structure-based drug design. Nature. 1996;384(6604 Suppl):23–26.

[15] Jorgensen WL. The Many Roles of Computation in Drug Discovery. Science. 2004;303(5665):1813 –1818.

[16] Jorgensen WL. Drug discovery: Pulled from a protein's embrace. Nature. 2010;466(7302):42–43.

[17] Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Berichte. 1894;27(3):2985–2993.

[18] Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis*. Proc Natl Acad Sci U S A. 1958;44(2):98–104.

[19] Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009;5(11):789–796.

[20] Ma B, Nussinov R. Enzyme dynamics point to stepwise conformational selection in catalysis. Curr Opin Chem Biol. 2010;14(5):652–659.

[21] Changeux J, Edelstein S. Conformational selection or induced fit? 50 years of debate resolved. F1000 Biology Reports. 2011;3:19.

[22] Hammes GG, Chang Y, Oas TG. Conformational selection or induced fit: A flux description of reaction mechanism. Proc Natl Acad Sci U S A. 2009;106(33):13737 –13741.

[23] Stein A, Rueda M, Panjkovich A, Orozco M, Aloy P. A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. Structure. 2011;19(6):881–889.

[24] Novoa EM, Pouplana LRd, Barril X, Orozco M. Ensemble Docking from Homology Models. J Chem Theory Comput. 2010;6(8):2547–2557.

[25] Huang S, Zou X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins. 2007;66(2):399–421.

[26] Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. J Med Chem. 2008;51(13):3878–3894.

[27] Giorgino T, Buch I, De Fabritiis G. Visualizing the Induced Binding of SH2-Phosphopeptide. J Chem Theory Comput. 2012;8(4):1171–1175.

[28] Morra G, Genoni A, Neves MAC, Merz J Kenneth M, Colombo G. Molecular recognition and drug-lead identification: what can molecular simulations tell us? Curr Med Chem. 2010;17(1):25–41.

[29] Ivetac A, McCammon JA. Molecular recognition in the case of flexible targets. Curr Pharm Des. 2011;17(17):1663–1671.

[30] Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, et al. Target flexibility: an emerging consideration in drug discovery and design. J Med Chem. 2008;51(20):6237–6255.

[31] Moore WJ. Physical chemistry. Prentice-Hall; 1972.

[32] General IJ. A Note on the Standard State's Binding Free Energy. J Chem Theory Comput. 2010;6(8):2520–2524.

[33] Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl. 2002;41(15):2644–2676.

[34] Chang C, Chen W, Gilson M. Ligand configurational entropy and protein binding. Proc Natl Acad Sci U S A. 2007;104(5):1534–1539.

[35] Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000;295(2):337–356.

[36] Talhout R, Villa A, Mark AE, Engberts JBFN. Understanding Binding Affinity: A Combined Isothermal Titration Calorimetry/Molecular Dynamics Study of the Binding of a Series of Hydrophobically Modified Benzamidinium Chloride Inhibitors to Trypsin. J Am Chem Soc. 2003;125(35):10570–10579.

[37] Olsson TS, Williams MA, Pitt WR, Ladbury JE. The Thermodynamics of Protein–Ligand Interaction and Solvation: Insights for Ligand Design. J Mol Biol. 2008;384(4):1002–1017.

[38] Copeland RA, Pompliano DL, Meek TD. Drug-target residence time and its implications for lead optimization. Nat Rev Drug Discov. 2006;5(9):730–739.

[39] Tummino PJ, Copeland RA. Residence time of receptor-ligand complexes and its effect on biological function. Biochemistry. 2008;47(20):5481–5492.

[40] Swinney DC. Biochemical mechanisms of drug action: what does it take for success? Nat Rev Drug Discov. 2004;3(9):801–808.

[41] Swinney DC. The role of binding kinetics in therapeutically useful drug action. Curr Opin Drug Discovery Dev. 2009;12(1):31–39.

[42] Swinney DC. Can Binding Kinetics Translate to a Clinically Differentiated Drug? From Theory to Practice. Lett Drug Des Discov. 2006;3(8):569–574.

[43] Schmidtke P, Luque FJ, Murray JB, Barril X. Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. J Am Chem Soc. 2011;133(46):18903–18910.

[44] Jung LS, Campbell CT, Chinowsky TM, Mar MN, Yee SS. Quantitative Interpretation of the Response of Surface Plasmon Resonance Sensors to Adsorbed Films. Langmuir. 1998;14(19):5636–5648.

[45] Mathews CK, Holde KEV, Ahern KG. Biochemistry. Benjamin Cummings; 2000.

[46] Nienhaus GU. Protein-ligand interactions: methods and applications. Humana Press; 2005.

[47] van Gunsteren WF, Dolenc J, Mark AE. Molecular simulation as an aid to experimentalists. Curr Opin Struct Biol. 2008;18(2):149–153.

[48] Leavitt S, Freire E. Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr Opin Struct Biol. 2001;11(5):560–566.

[49] Wiseman T, Williston S, Brandts JF, Lin LN. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem. 1989;179(1):131–137.

[50] Rich RL, Hoth LR, Geoghegan KF, Brown TA, LeMotte PK, Simons SP, et al. Kinetic analysis of estrogen receptor/ligand interactions. Proc Natl Acad Sci U S A. 2002;99(13):8562–8567.

[51] Frey BL, Jordan CE, Kornguth S, Corn RM. Control of the Specific Adsorption of Proteins onto Gold Surfaces with Poly(L-lysine) Monolayers. Anal Chem. 1995;67(24):4452–4457.

[52] Li S, Schmitz KR, Jeffrey PD, Wiltzius JJW, Kussie P, Ferguson KM. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell. 2005;7(4):301–311.

[53] Heaton RJ, Peterson AW, Georgiadis RM. Electrostatic surface plasmon resonance: direct electric field-induced hybridization and denaturation in monolayer nucleic acid films and label-free discrimination of base mismatches. Proc Natl Acad Sci U S A. 2001;98(7):3701–3704.

[54] Brockman JM, Frutos AG, Corn RM. A Multistep Chemical Modification Procedure To Create DNA Arrays on Gold Surfaces for the Study of ProteinDNA Interactions with Surface Plasmon Resonance Imaging. J Am Chem Soc. 1999;121(35):8044–8051.

[55] Brockman JM, Nelson BP, Corn RM. Surface plasmon resonance imaging measurements of ultrathin organic films. Annu Rev Phys Chem. 2000;51:41–63.

[56] Haake HM, Schütz A, Gauglitz G. Label-free detection of biomolecular interaction by optical sensors. Fresen J Anal Chem. 2000;366(6-7):576–585.

[57] Karlsson R, Ståhlberg R. Surface plasmon resonance detection and multispot sensing for direct monitoring of interactions involving low-molecular-weight analytes and for determination of low affinities. Anal Biochem. 1995;228(2):274–280.

[58] Haes AJ, Van Duyne RP. A Nanoscale Optical Biosensor: Sensitivity and Selectivity of an Approach Based on the Localized Surface Plasmon Resonance Spectroscopy of Triangular Silver Nanoparticles. J Am Chem Soc. 2002;124(35):10596–10604.

[59] Papalia GA, Baer M, Luehrsen K, Nordin H, Flynn P, Myszka DG. High-resolution characterization of antibody fragment/antigen interactions using Biacore T100. Anal Biochem. 2006;359(1):112–119.

[60] Rich RL, Myszka DG. Kinetic analysis and fragment screening with Fujifilm AP-3000. Anal Biochem. 2010;402(2):170–178.

[61] Ogiso H, Ishitani R, Nureki O, Fukai S, Yamanaka M, Kim J, et al. Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell. 2002;110(6):775–787.

[62] Bourgeois D, Royant A. Advances in kinetic protein crystallography. Curr Opin Struct Biol. 2005;15(5):538–547.

[63] Schlichting I, Chu K. Trapping intermediates in the crystal: ligand binding to myoglobin. Curr Opin Struct Biol. 2000;10(6):744–752.

[64] Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007;450(7172):964–972.

[65] Karplus M, Petsko GA. Molecular dynamics simulations in biology. Nature. 1990;347(6294):631–639.

[66] Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An Automated Protein Homology-Modeling Server. Nucleic Acids Res. 2003;31(13):3381–3385.

[67] Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science. 2010;330(6002):341–346.

[68] Boyce SE, Mobley DL, Rocklin GJ, Graves AP, Dill KA, Shoichet BK. Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. J Mol Biol. 2009;394(4):747–763.

[69] Mobley DL. Let's get honest about sampling. J Comput-Aided Mol Des. 2012;26(1):93–95.

[70] Jayachandran G, Shirts MR, Park S, Pande VS. Parallelized-over-parts computation of absolute binding free energy with docking and molecular dynamics. J Chem Phys. 2006;125(8):084901.

[71] Buch I, Sadiq SK, De Fabritiis G. Optimized Potential of Mean Force Calculations for Standard Binding Free Energies. J Chem Theory Comput. 2011;7(6):1765–1772.

[72] Gallicchio E, Levy RM. Advances in all atom sampling methods for modeling protein-ligand binding affinities. Curr Opin Struct Biol. 2011;21(2):161–166.

[73] Luccarelli J, Michel J, Tirado-Rives J, Jorgensen WL. Effects of Water Placement on Predictions of Binding Affinities for p38 MAP Kinase Inhibitors. J Chem Theory Comput. 2010;6(12):3850–3856.

[74] Murdock SE, Tai K, Ng MH, Johnston S, Wu B, Fangohr H, et al. Quality Assurance for Biomolecular Simulations. J Chem Theory Comput. 2006;2(6):1477–1481.

[75] Brannigan G, LeBard DN, Hénin J, Eckenhoff RG, Klein ML. Multiple binding sites for the general anesthetic isoflurane identified in the nicotinic acetylcholine receptor transmembrane domain. Proc Natl Acad Sci U S A. 2010;107(32):14122.

[76] Vanni S, Neri M, Tavernelli I, Rothlisberger U. Predicting Novel Binding Modes of Agonists to Adrenergic Receptors Using All-Atom Molecular Dynamics Simulations. PLoS Comput Biol. 2011;7(1):e1001053.

[77] Enkavi G, Tajkhorshid E. Simulation of Spontaneous Substrate Binding Revealing the Binding Pathway and Mechanism and Initial Conformational Response of GlpT. Biochemistry. 2010;49(6):1105–1114.

[78] Borhani DW, Shaw DE. The future of molecular dynamics simulations in drug discovery. J Comput-Aided Mol Des. 2012;26(1):15–26.

[79] Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, et al. Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM. 2008;51(7):91–97.

[80] Harvey MJ, Giupponi G, Fabritiis GD. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J Chem Theory Comput. 2009;5(6):1632–1639.

[81] Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. J Chem Inf Model. 2010;50(3):397–403.

[82] Beberg AL, Ensign DL, Jayachandran G, Khaliq S, Pande VS. Folding@home: Lessons from eight years of volunteer distributed computing. In: Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing. IPDPS '09. Washington, DC, USA: IEEE Computer Society; 2009. p. 1–8.

[83] Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011;9(1):71.

[84] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins. 2006;65(3):712–725.

[85] Best RB, Buchete N, Hummer G. Are current molecular dynamics force fields too helical? Biophys J. 2008;95(1):L07–09.

[86] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010;78(8):1950–1958.

[87] Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J. 2011;100(9):L47–49.

[88] Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic Validation of Protein Force Fields against Experimental Data. PLoS ONE. 2012;7(2):e32131.

[89] Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc. 1988;110(6):1657–1666.

[90] Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides†. J Phys Chem B. 2001;105(28):6474–6487.

[91] Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem. 2003;24(16):1999–2012.

[92] MacKerell, Bashford D, Bellott, Dunbrack, Evanseck JD, Field MJ, et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. J Phys Chem B. 1998;102(18):3586–3616.

[93] Mackerell J Alexander D, Feig M, Brooks r Charles L. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem. 2004;25(11):1400–1415.

[94] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004;25(9):1157–1174.

[95] Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem. 2002;23(16):1623–1641.

[96] Jiao D, Golubkov PA, Darden TA, Ren P. Calculation of Protein–ligand Binding Free Energy by Using a Polarizable Potential. Proc Natl Acad Sci U S A. 2008;105(17):6290–6295.

[97] Pospisil P, Ballmer P, Scapozza L, Folkers G. Tautomerism in computer-aided drug design. J Recept Sig Transd. 2003;23(4):361–371.

[98] Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins. 2006;65(1):15–26.

[99] Graves AP, Shivakumar DM, Boyce SE, Jacobson MP, Case DA, Shoichet BK. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. J Mol Biol. 2008;377(3):914–934.

[100] Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, et al. Development and validation of a modular, extensible docking program: DOCK 5. J Comput-Aided Mol Des. 2006;20(10-11):601–619.

[101] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004;47(7):1739–1749.

[102] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–461.

[103] Wang J, Dixon R, Kollman PA. Ranking ligand binding affinities with avidin: a molecular dynamics-based interaction energy study. Proteins. 1999;34(1):69–81.

[104] Aqvist J, Medina C, Samuelsson JE. A new method for predicting binding affinity in computer-aided drug design. Protein Eng. 1994;7(3):385–391.

[105] Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. Continuum Solvent Studies of the Stability of DNA, RNA, and PhosphoramidateDNA Helices. J Am Chem Soc. 1998;120(37):9401–9409.

[106] Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Acc Chem Res. 2000;33(12):889–897.

[107] Huo S, Wang J, Cieplak P, Kollman PA, Kuntz ID. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. J Med Chem. 2002;45(7):1412–1419.

[108] Rizzo RC, Toba S, Kuntz ID. A Molecular Basis for the Selectivity of Thiadiazole Urea Inhibitors with Stromelysin-1 and Gelatinase-A from Generalized Born Molecular Dynamics Simulations. J Med Chem. 2004;47(12):3065–3074.

[109] Stoica I, Sadiq SK, Coveney PV. Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. J Am Chem Soc. 2008;130(8):2639–2648.

[110] Lu N, Singh JK, Kofke DA. Appropriate methods to combine forward and reverse free-energy perturbation averages. J Chem Phys. 2003;118(7):2977–2984.

[111] Price DJ, Jorgensen WL. Improved convergence of binding affinities with free energy perturbation: application to nonpeptide ligands with pp60src SH2 domain. J Comput-Aided Mol Des. 2001;15(8):681–695.

[112] Reddy MR, Erion MD. Calculation of Relative Binding Free Energy Differences for Fructose 1,6-Bisphosphatase Inhibitors Using the

Thermodynamic Cycle Perturbation Approach. J Am Chem Soc. 2001;123(26):6246–6252.

[113] Shirts MR, Pande VS. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. J Chem Phys. 2005;122(14):144107–144107–16.

[114] Fowler PW, Jha S, Coveney PV. Grid-Based Steered Thermodynamic Integration Accelerates the Calculation of Binding Free Energies. Philos T Roy Soc A. 2005;363(1833):1999–2015.

[115] Wan S, Coveney PV, Flower DR. Peptide Recognition by the T Cell Receptor: Comparison of Binding Free Energies from Thermodynamic Integration, Poisson–Boltzmann and Linear Interaction Energy Approximations. Philos T Roy Soc A. 2005;363(1833):2037–2053.

[116] Mobley DL, Chodera JD, Dill KA. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. J Chem Phys. 2006;125(8):084902.

[117] Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. J Am Chem Soc. 2005;127(8):2600–2607.

[118] Fidelak J, Juraszek J, Branduardi D, Bianciotto M, Gervasio FL. Free-Energy-Based Methods for Binding Profile Determination in a Congeneric Series of CDK2 Inhibitors. J Phys Chem B. 2010;114(29):9516–9524.

[119] Selent J, Sanz F, Pastor M, De Fabritiis G. Induced Effects of Sodium Ions on Dopaminergic G-Protein Coupled Receptors. PLoS Comput Biol. 2010;6(8):e1000884.

[120] De Fabritiis G, Coveney PV, Villà-Freixa J. Energetics of K+ permeability through Gramicidin A by forward-reverse steered molecular dynamics. Proteins. 2008;73(1):185–194.

[121] Giorgino T, De Fabritiis G. A High-Throughput Steered Molecular Dynamics Study on the Free Energy Profile of Ion Permeation through Gramicidin A. J Chem Theory Comput. 2011;7(6):1943–1950.

[122] Colizzi F, Perozzo R, Scapozza L, Recanatini M, Cavalli A. Single-Molecule Pulling Simulations Can Discern Active from Inactive Enzyme Inhibitors. J Am Chem Soc. 2010;132(21):7361–7371.

[123] Roux B. The calculation of the potential of mean force using computer simulations. Comput Phys Commun. 1995;91(1-3):275–282.

[124] Woo H, Roux B. Calculation of absolute protein–ligand binding free energy from computer simulations. Proc Natl Acad Sci U S A. 2005;102(19):6825–6830.

[125] Shirts MR, Pitera JW, Swope WC, Pande VS. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. J Chem Phys. 2003;119(11):5740–5761.

[126] Michel J, Essex JW. Hit identification and binding mode predictions by rigorous free energy simulations. J Med Chem. 2008;51(21):6654–6664.

[127] Merz KM. Limits of Free Energy Computation for ProteinLigand Interactions. J Chem Theory Comput. 2010;6(5):1769–1776.

[128] Mobley DL, Dill KA. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". Structure. 2009;17(4):489–498.

[129] Malmstrom RD, Watowich SJ. Using Free Energy of Binding Calculations To Improve the Accuracy of Virtual Screening Predictions. J Chem Inf Model. 2011;51(7):1648–1655.

[130] Benson ML, Faver JC, Ucisik MN, Dashti DS, Zheng Z, Merz KM. Prediction of trypsin/molecular fragment binding affinities by free energy decomposition and empirical scores. J Comput-Aided Mol Des. 2012;in press.

[131] Michel J, Essex JW. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput-Aided Mol Des. 2010;24(8):639–658.

[132] Krohn KA, Link JM. Interpreting enzyme and receptor kinetics: keeping it simple, but not too simple. Nucl Med Biol. 2003;30(8):819–826.

[133] Boresch S, Tettinger F, Leitgeb M, Karplus M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. J Phys Chem B. 2003;107(35):9535–9551.

[134] Doudou S, Burton NA, Henchman RH. Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. J Chem Theory Comput. 2009;5(4):909–918.

[135] Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci U S A. 2011 Jun;108(25):10184–10189.

[136] Torrie G, Valleau J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J Comput Phys. 1977;23(2):187–199.

[137] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. THE weighted histogram analysis method for freeenergy calculations on biomolecules. I. The method. J Comput Chem. 1992;13(8):1011–1021.

[138] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Multidimensional free-energy calculations using the weighted histogram analysis method. J Comput Chem. 1995;16(11):1339–1350.

[139] Juraszek J, Bolhuis PG. Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. Biophys J. 2008;95(9):4246–4257.

[140] Wang J, Deng Y, Roux B. Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. Biophys J. 2006;91(8):2798–2814.

[141] Deng Y, Roux B. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. J Chem Phys. 2008;128(11):115103.

[142] Singhal N, Snow CD, Pande VS. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. J Chem Phys. 2004;121(1):415–425.

[143] Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J Chem Phys. 2007;126(15):155101.

[144] Noé F, Fischer S. Transition networks for modeling the kinetics of conformational change in macromolecules. Curr Opin Struct Biol. 2008;18(2):154–162.

[145] Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010;52(1):99–105.

[146] Prinz J, Wu H, Sarich M, Keller B, Senne M, Held M, et al. Markov models of molecular kinetics: Generation and validation. J Chem Phys. 2011;134(17):174105.

[147] Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci U S A. 2009;106(45):19011 –19016.

[148] Voelz VA, Bowman GR, Beauchamp K, Pande VS. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). J Am Chem Soc. 2010;132(5):1526–1528.

[149] Radford IH, Fersht AR, Settanni G. Combination of Markov State Models and Kinetic Networks for the Analysis of Molecular Dynamics Simulations of Peptide Folding. J Phys Chem B. 2011;115(22):7459–7471.

[150] Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. J Am Chem Soc. 2011;133(45):18413–18419.

[151] Held M, Metzner P, Prinz J, Noé F. Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations. Biophys J. 2011;100(3):701–710.

[152] Silva D, Bowman GR, Sosa-Peinado A, Huang X. A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. PLoS Comput Biol. 2011;7(5):e1002054.

[153] Kusch J, Thon S, Schulz E, Biskup C, Nache V, Zimmer T, et al. How subunits cooperate in cAMP-induced activation of homotetrameric HCN2 channels. Nat Chem Biol. 2012;8(2):162–169.

[154] Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proc Natl Acad Sci U S A. 2004;101(41):14766–14770.

[155] Shirts M, Pande VS. Screen Savers of the World Unite! Science. 2000;290(5498):1903–1904.

[156] Huang D, Caflisch A. The Free Energy Landscape of Small Molecule Unbinding. PLoS Comput Biol. 2011;7(2):e1002002.

[157] Moore GE. Lithography and the future of Moore's law. vol. 2437. SPIE; 1995. p. 2–17.

[158] Giupponi G, Harvey M, De Fabritiis G. The impact of accelerator processors for high-throughput molecular modeling and simulation. Drug Discov Today. 2008;13(23-24):1052–1058.

[159] Harvey MJ, De Fabritiis G. An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware. J Chem Theory Comput. 2009;5(9):2371–2377.

[160] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26(16):1781–1802.

[161] Case DA, Cheatham r Thomas E, Darden T, Gohlke H, Luo R, Merz J Kenneth M, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26(16):1668–1688.

[162] Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008;4(3):435–447.

[163] Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the 2006 ACM/IEEE conference on Supercomputing. Tampa, Florida: ACM; 2006. p. 84.

[164] Anderson DP. Boinc: A system for public-resource computing and storage. In: 5th IEEE/ACM International Workshop on Grid Computing; 2004. p. 4–10.

[165] Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D. SETI@home: an experiment in public-resource computing. Commun ACM. 2002;45(11):56–61.

[166] TOP 500; Accessed on March 2012. Available from: http://www.top500.org/.

[167] Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovic Z, et al. Algorithm discovery by protein folding game players. Proc Natl Acad Sci U S A. 2011 Nov;108(47):18949–18953.

[168] GPUGRID.net;. Available from: http://www.gpugrid.net.

[169] Giorgino T, Harvey M, de Fabritiis G. Distributed computing as a virtual supercomputer: Tools to run and manage large-scale BOINC simulations. Comput Phys Commun. 2010;181(8):1402–1409.

[170] AllProjectStats; Accessed on March 2012. Available from: http://www.allprojectstats.com/.

[171] World Community Grid;. Available from: http://www.worldcommunitygrid.org/.

[172] PrimeGrid;. Available from: http://www.primegrid.com/.

[173] Yoyo@Home;. Available from: http://www.rechenkraft.net/yoyo.

[174] Wikipedia contributors. Gamification. Wikimedia Foundation, Inc.; 2012. Page Version ID: 486852138.

[175] Radoff J. Game on: energize your business with social media games. New York: Wiley Publishing, Inc.; 2011.

[176] Sadowski I, Stone JC, Pawson T. A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. Mol Cell Biol. 1986;6(12):4396–4408.

[177] Pawson T, Gish GD. SH2 and SH3 domains: from structure to function. Cell. 1992;71(3):359–362.

[178] Lowenstein EJ, Daly RJ, Batzer AG, Li W, Margolis B, Lammers R, et al. The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling. Cell. 1992;70(3):431–442.

[179] Songyang Z, Shoelson SE, Chaudhuri M, Gish G, Pawson T, Haser WG, et al. SH2 domains recognize specific phosphopeptide sequences. Cell. 1993;72(5):767–778.

[180] Tong L. Crystal Structures of the Human p56lckSH2 Domain in Complex with Two Short Phosphotyrosyl Peptides at 1.0 Å and 1.8 Å Resolution. J Mol Biol. 1996;256(3):601–610.

[181] Bradshaw J, Mitaxov V, Waksman G. Mutational investigation of the specificity determining region of the src SH2 domain1. J Mol Biol. 2000;299(2):523–537.

[182] Eck MJ, Shoelson SE, Harrison SC. Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. Nature. 1993;362(6415):87–91.

[183] Carter P, Wells JA. Dissecting the catalytic triad of a serine protease. Nature. 1988;332(6164):564–568.

[184] Polgár L. The catalytic triad of serine peptidases. Cell Mol Life Sci. 2005;62(19):2161–2172.

[185] Huber R, Kukla D, Bode W, Schwager P, Bartels K, Deisenhofer J, et al. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II. Crystallographic refinement at 1.9 A resolution. J Mol Biol. 1974;89(1):73–101.

[186] Rühlmann A, Kukla D, Schwager P, Bartels K, Huber R. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. Crystal structure determination and stereochemistry of the contact region. J Mol Biol. 1973;77(3):417–436.

[187] Bode W, Schwager P. The refined crystal structure of bovine -trypsin at 1·8 Å resolutionII. Crystallographic refinement, calcium binding site, benzamidine binding site and active site at pH 7·0. J Mol Biol. 1975;98(4):693–717.

[188] Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr B. 1983;39(4):480–490.

[189] Holbro T, Hynes NE. ErbB receptors: directing key signaling networks throughout life. Annu Rev Pharmacol Toxicol. 2004;44:195–217.

[190] Arteaga C. Targeting HER1/EGFR: a molecular approach to cancer therapy. Semin Oncol. 2003;30(3 Suppl 7):3–14.

[191] Arteaga CL. Epidermal Growth Factor Receptor Dependence in Human Tumors: More Than Just Expression? Oncologist. 2002;7(90004):31–39.

[192] Boerner JL, Danielsen AJ, Lovejoy CA, Wang Z, Juneja SC, Faupel-Badger JM, et al. Grb2 regulation of the actin-based cytoskeleton is required for ligand-independent EGF receptor-mediated oncogenesis. Oncogene. 2003;22(43):6679–6689.

[193] Normanno N, Bianco C, De Luca A, Salomon DS. The role of EGF-related peptides in tumor growth. Front Biosci. 2001;6:D685–707.

[194] Gullick WJ. Prevalence of aberrant expression of the epidermal growth factor receptor in human cancers. Br Med Bull. 1991;47(1):87–98.

[195] Klijn JG, Berns PM, Schmitz PI, Foekens JA. The clinical significance of epidermal growth factor receptor (EGF-R) in human breast cancer: a review on 5232 patients. Endocr Rev. 1992;13(1):3–17.

[196] Sainsbury JR, Malcolm AJ, Appleton DR, Farndon JR, Harris AL. Presence of epidermal growth factor receptor as an indicator of poor prognosis in patients with breast cancer. Am J Clin Path. 1985;38(11):1225–1228.

[197] Salomon DS, Brandt R, Ciardiello F, Normanno N. Epidermal growth factor-related peptides and their receptors in human malignancies. Crit Rev Oncol Hematol. 1995;19(3):183–232.

[198] Yang XD, Jia XC, Corvalan JR, Wang P, Davis CG. Development of ABX-EGF, a fully human anti-EGF receptor monoclonal antibody, for cancer therapy. Crit Rev Oncol Hematol. 2001;38(1):17–23.

[199] Montagut C, Dalmases A, Bellosillo B, Crespo M, Pairet S, Iglesias M, et al. Identification of a mutation in the extracellular domain of the Epidermal Growth Factor Receptor conferring cetuximab resistance in colorectal cancer. Nat Med. 2012;18(2):221–223.

[200] De Fabritiis G. Performance of the Cell processor for biomolecular simulations. Comput Phys Commun. 2007;176(11-12):660 –664.

[201] Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins. 2007;69 Suppl 8:118–128.

[202] BoincStats; Accessed on March 2012. Available from: http://www.boincstats.com/.

[203] Anderson DP. Public Computing: Reconnecting People to Science; 2003. http://boinc.berkeley.edu/madrid.html.

[204] Sadiq SK, De Fabritiis G. Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. Proteins. 2010;78(14):2873–2885.

[205] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science. 2011;334(6055):517–520.

[206] Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan Y, et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc Natl Acad Sci U S A. 2011;108(32):13118 –13123.

[207] Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How Does a Drug Molecule Find Its Target Binding Site? J Am Chem Soc. 2011;133(24):9181–9183.

[208] Kollman P. Free energy calculations: Applications to chemical and biochemical phenomena. Chem Rev. 1993;93(7):2395–2417.

[209] Gilson MK, Zhou H. Calculation of Protein-Ligand Binding Affinities*. Annu Rev Biophys Biomol Struct. 2007;36(1):21–42.

[210] Deng Y, Roux B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. J Phys Chem B. 2009;113(8):2234–2246.

[211] Becker OM, Jr ADM, Roux B, Watanabe M. Computational Biochemistry and Biophysics. CRC Press; 2001.

[212] Barducci A, Bonomi M, Parrinello M. Metadynamics. WIREs Comput Mol Sci. 2011;1(5):826–843.

[213] Laio A, Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys. 2008;71(12):126601.

[214] Saladino G, Gauthier L, Bianciotto M, Gervasio FL. Assessing the Performance of Metadynamics and Path Variables in Predicting the Binding Free Energies of p38 Inhibitors. J Chem Theory Comput. 2012;8(4):1165–1170.

[215] Lee TR, Lawrence DS. SH2-Directed Ligands of the Lck Tyrosine Kinase. J Med Chem. 2000;43(6):1173–1179.

[216] Mobley DL, Chodera JD, Dill KA. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. J Chem Theory Comput. 2007;3(4):1231–1235.

[217] General IJ, Dragomirova R, Meirovitch H. Calculation of the Absolute Free Energy of Binding and Related Entropies with the HSMD-TI Method: The FKBP12-L8 Complex. J Chem Theory Comput. 2011;7(12):4196–4207.

[218] Patel JS, Branduardi D, Masetti M, Rocchia W, Cavalli A. Insights into Ligand–Protein Binding from Local Mechanical Response. J Chem Theory Comput. 2011;7(10):3368–3378.

[219] Lapelosa M, Gallicchio E, Levy RM. Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations. J Chem Theory Comput. 2012;8(1):47–60.

[220] Li S, Kussie P, Ferguson KM. Structural Basis for EGF Receptor Inhibition by the Therapeutic Antibody IMC-11F8. Structure. 2008;16(2):216–227.

[221] Yun C, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong K, et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc Natl Acad Sci U S A. 2008;105(6):2070–2075.

[222] Bowles JA, Wang S, Link BK, Allan B, Beuerlein G, Campbell M, et al. Anti-CD20 monoclonal antibody with enhanced affinity for CD16 activates NK cells at lower concentrations and more effectively than rituximab. Blood. 2006;108(8):2648–2654.

[223] Li B, Zhao L, Guo H, Wang C, Zhang X, Wu L, et al. Characterization of a rituximab variant with potent antitumor activity against rituximab-resistant B-cell lymphoma. Blood. 2009;114(24):5007–5015.

[224] Ahmad M, Gu W, Helms V. Mechanism of Fast Peptide Recognition by SH3 Domains13. Angew Chem Int Ed Engl. 2008;47(40):7626–7630.

[225] Wu C, Biancalana M, Koide S, Shea J. Binding Modes of Thioflavin-T to the Single-Layer [beta]-Sheet of the Peptide Self-Assembly Mimics. J Mol Biol. 2009;394(4):627–633.

[226] Lange OF, Grubmüller H. Full correlation analysis of conformational protein dynamics. Proteins. 2008;70(4):1294–1312.

[227] Tribello GA, Ceriotti M, Parrinello M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. Proc Natl Acad Sci U S A. 2012;109(14):5196–5201.

[228] Metzner P, Schütte C, Vanden-Eijnden E. Transition Path Theory for Markov Jump Processes. Multiscale Model Sim. 2009;7(3):1192.

[229] Held M, Noé F. Calculating kinetics and pathways of protein–ligand association. Eur J Cell Biol. 2012;91(4):357–364.

[230] Söderhjelm P, Tribello GA, Parrinello M. Locating Binding Poses in Protein-Ligand Systems Using Reconnaissance Metadynamics. Proc Natl Acad Sci U S A. 2012;109(14):5170–5175.

[231] Burkhardt A, Warmer M, Panneerselvam S, Wagner A, Zouni A, Glöckner C, et al. Fast high-pressure freezing of protein crystals in their mother liquor. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2012;68(4):495–500.

[232] Lu H, Tonge PJ. Drug-target residence time: critical information for lead optimization. Curr Opin Chem Biol. 2010;14(4):467–474.

[233] Bowman GR, Ensign DL, Pande VS. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. J Chem Theory Comput. 2010;6(3):787–794.

[234] Hajduk PJ, Greer J. A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov. 2007;6(3):211–219.

[235] Gill AL, Frederickson M, Cleasby A, Woodhead SJ, Carr MG, Woodhead AJ, et al. Identification of novel p38alpha MAP kinase inhibitors using fragment-based lead generation. J Med Chem. 2005;48(2):414–426.

[236] Agamennone M, Cesari L, Lalli D, Turlizzi E, DelConte R, Turano P, et al. Fragmenting the S100B-p53 Interaction: Combined Virtual/Biophysical Screening Approaches to Identify Ligands. ChemMedChem. 2010;5(3):428–435.