Mapping Biophysics through enhanced Monte Carlo techniques

by

Israel Cabeza de Vaca Lopez

A dissertation presented to

The department of Applied Physics at Universitat Politècnica de Catalunya

In partial fulfilment of the requirements for the degree of Doctor per la Universitat Politècnica de Catalunya in the subject of Computational and Applied Physics

Computational and Applied Physics Program

Advisor: Prof. Victor Guallar Tasies
Tutor: Prof. Manel Canales Gabriel

Universitat Politècnica de Catalunya
Life Science Department at Barcelona Supercomputing Center

November 2015

Dedico esta tesis a mis padres José María y Asunción por su incansable esfuerzo, apoyo y dedicación.

Sin vuestra ayuda esta tesis no habria sido posible.

“Work until your idols become your rivals”

Drake

Table of Contents

ABSTRACT

Mapping Biophysics through enhanced Monte Carlo techniques

This thesis is focused on the development, improvement and application of Monte Carlo algorithms to study molecular interactions at atomistic level using classical molecular mechanics. In this thesis, we have worked on three main parts: Protein/DNA-ligand interactions, steering of proteins and multiscale approach for protein-protein docking, using as a frame our in-house algorithm PELE. We proposed a bound model in close collaboration with experimental groups for porphyrin and bisphophate sugar ligands. Later, we reproduced similar DNA fluctuations than the well-established molecular dynamics method for different representative DNA fragments. Moreover, we studied DNA-Cisplatin drugs interactions, and we evaluated the binding free energies of these compounds with excellent agreement with molecular dynamics results. Furthermore, we measured the force extension profile of the ubiquitin and azurin proteins stretching the proteins. Lastly, we applied a multiscale approach based on a coarse-grained model and all-atom refinement to generate and score protein-protein conformations in three systems.

Keywords: Monte Carlo, Molecular Dynamics, PELE, induced fit docking, Markov States Models, Coarse grained, Protein/DNA-ligand interactions, intercalators.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to everyone that directly or indirectly have contributed to the development of my doctoral thesis.

I will start the acknowledgments with my thesis director Professor Victor Guallar Tasies. He has been always giving me support during my Ph.D. thesis even when the good results were not coming. I have learned a lot from thousands of scientific discussions that we had during our meeting. When I arrived at his group I had no clue about how to develop a scientific project. Now, thanks to Victor, I have learned how to think and reason in a scientific way. He taught me with his enthusiastic and motivational personality how to present the scientific work at international group talks and how to publish the work in peer-reviewed journals. For all these reason and many others, since the very beginning of my thesis, he became a mentor and a reference to me.

I want to express my gratitude to my UPC tutor Manel Canales Gabriel for his suggestions and help to focus the thesis and his time helping me to prepare the thesis. Also, special thanks to the external collaborators for the different projects developed during my thesis: Juan Fernandez-Recio, Pau Gorostiza, Marina Ines Giannotti, James Valdes, Maria Victoria Cubellis, Roman Sobotka, Jana Kopecna, Milagros Medina, Jaime Rubio-Martinez and many others. Moreover, I would like to thank the UPC doctoral secretary Ana Maria Ortega for her useful help with the Ph.D. bureaucracy. Special thanks to the Barcelona Supercomputing Center (BSC) for hosting and offer me significant computational resources needed for my thesis. Also, thanks to the BSC Severo Ochoa Mobility grant provided to help me to stay at YALE University during one part of my thesis.

I also would like to thank my colleagues. During these last years, I have had the opportunity to meet and collaborate with many persons, and I have learned something from each of them. Some of them have also become friends another just colleagues, but I am pretty sure that a bit of each of them will remain as part of my personality. Thanks goes to Suwipa, Jorge, Pedro, Ali, Frank, Ken, Ben, Diego, Max, Victor Gil, Armin, Fatima, Marina, Ferrán, Ryoji, Dani, Martin, Marcelo, Manuel, Emmanuelle and other BSC members. I will always remember the great moments at the Nuria retreats, beers after work in FIB bar and group dinners.

My colleague and girlfriend Sandra deserves another section. I will always appreciate her support during my Ph.D. Working until late when no one was in the office, long weekends in the office and the best moment of the day when we arrived exhausted to our home after a hard workday. She always helps me to be focused on my thesis, and she never complained about the enormous amount of hours required by my projects. I have no words to express my gratitude.

I would like to thanks to my friends Ricardo, Antonio, Dani and David for his support and great moments since I met them during my Bachelor in Physics at the University of Barcelona. I have been so lucky having them as good friends. I will always remember the sailing trips with Dani and David around Costa Brava and Menorca Island, the Antonio's discussions in the UB physics faculty bar about any weird technological topic and the Ricardo's adventures and misfortunes in New York and New Haven when we meet there.

Special thanks to my friends Dani and Marta Aroa and them puppy Atrus for the priceless moments on the sailboat, having dinner or lunch, hiking in the Catalonian mountains and a lot of other lovely moments.

Very special thanks go to my whole Family and especially to my dad Jose Maria and my mom Chon who always supported and helped me during the school, high school, Bachelor, Masters and the Ph.D. They always believed in me and my skills, and I will never be able to express with words my appreciation for this priceless support.

The last gratitude but equally important goes to Prof. Bill Jorgensen and his group. I had the opportunity to learn and work with them during more than three months developing one small part of this thesis. Special thanks go to Paty and Julian, who make me feel part of the group from the first moment. I will always be in debt with them for the scientific and personal support during my stay, and I will always remember the BBQs and other special moments shared with them. Thanks to my office mates John and Danny for the great beer moments after work and their help in the lab with my project. Jose and Marga also deserve a special mention because I was so lucky to meet them in the group. They helped me a lot taking me to many places by car and introducing me nice drinking buddies in New Haven from other YALE departments. Also, thanks to Pawel, Luca, Ana, Vinay, Won-Gil, Kara, Cherry, Michael, Jona, Leela and Cindy for the great lunch time conversations.

PREFACE

This thesis collects the majority of my projects developed during my Ph.D. in the Electronic and Atomic Protein Modelling group in the Life Science department at Barcelona Supercomputing Center. The main motivation of this thesis has been the improvement and development of new and efficient Monte Carlo algorithms to accelerate drug discovery and study biophysical interactions. Increasing computer power and reduction of associated costs makes computer simulations an excellent alternative to speed up the biophysical research saving money and time respect to traditional experiments. For this reason, pharmaceutical companies and research institutes are investing money in computational resources and qualified persons. For instance, last years, computational departments in pharmaceutical companies have become the first step in drug discovery selecting sets of potentially suitable candidates, and it has increased the importance of accurate initial computational predictions.

This thesis is focused on the study of molecular interactions at the atomistic detail and is divided into one introductory chapter and four chapters referencing different problems and methodological approaches. All of them are focused on the development and improvement of computational Monte Carlo algorithms to study, in an efficient manner, the behavior of these systems at a classical molecular mechanics level. The four biophysical problems studied in this thesis are: induced fit docking between protein-ligand and between DNA-ligand to understand the binding mechanism, protein stretching response, and generation/scoring of protein-protein docking poses.

The manuscript is organized as follows: First chapter corresponds to the state of the art in computational methods to study biophysical interactions, which is the starting point of this thesis. Our in-house PELE algorithm and the main standard methods such as molecular dynamics will be explained in detail. Chapter two is focused on the main PELE modifications to add new features, such as the addition of a new force field, implicit solvent and an anisotropic network specific for DNA simulation studies. We will study, compare and validate the conformations generated by six representative DNA fragments with the new PELE features using molecular dynamics as a reference.

Chapter three is devoted to applying the new methods implemented and tested in PELE to study protein-ligand interactions and DNA-ligand interactions using four systems. First, we

will study the porphyrin binding to Gun4 protein combining PELE and molecular dynamics simulations. Besides, we will provide a docking pose that will be corroborated by a new crystal structure published during the revision process of the submitted study showing the accuracy of our predictions. In the second project, we will use our improved version of PELE to generate the first structural model of an alpha glucose 1,6-bisphosphate substrate bound to the human Phosphomannomutase 2 demonstrating that this ligand can adopt two low-energy orientations. The third project will be the study of DNA-ligand interactions for three cisplatin drugs where we will evaluate the binding free energy using Markov state models. We will show excellent results respect another free energy methods studied with molecular dynamics. The last project will be the study of the daunomycin DNA intercalator where we will simulate and study the binding process with PELE.

Chapter four is focused on the computational study of force extension profiles during the protein unfolding. We will add a dynamic harmonic constraint following a similar procedure applied in steered molecular dynamics to our Monte Carlo approach to fix or pull some selected atoms forcing the protein unfolding in a defined direction. We will implement and compare with steered molecular dynamics this technique with Ubiquitin and Azurin proteins. We will compare the force extension profile of Azurin holo and apo (with a coordinated copper or without, respectively) pulling different surface residues to obtain a similar distribution of rupture length of the atomic force microscopy experiments developed in collaboration with Pau Gorostiza and Marina Gianotti at Institut de Bioenginyeria de Catalunya (IBEC). Moreover, we will add this feature to a well-known algorithm called MCPRO from William Jorgensen's group at YALE University to evaluate the free energy associated to the unfolding of the deca-alanine system.

Chapter five corresponds to the introduction of a multiscale approach to study protein-protein docking. A coarse-grained model will be combined with a Monte Carlo exploration reducing the degrees of freedom to generate thousands of protein-protein poses in a quick way. Poses produced by this procedure will be refined and ranked through a protonation, hydrogen bond optimization, and minimization protocol at the all-atom representation to identify the best poses. I will present two test cases where this procedure has been applied showing a good accuracy in the predictions: tryptogalinin and ferredoxin/flavodoxin systems.

Finally, the last chapter is a general conclusion of the thesis focus on the objective, and the results obtained.

List of figures

List of tables

Abbreviations list

ns	Nanoseconds
M	Molar (mol/dm^3)
MD	Molecular dynamics
MC	Monte Carlo
CG	Coarse grained
MSM	Markov State Models
NA	Nucleic acid
DNA	Deoxyribonucleic acid
PDB	Protein data bank
NMR	Nuclear magnetic resonance
PB	Poisson-Boltzmann
GB	Generalized Born
GBSA	Generalized Born surface area
SGB	Surface generalized Born
SGBNP	Surface generalized Born with non polar
VDGBNP	Variable dielectric generalized Born with non polar
SMD	Steered molecular dynamics
PELE	Protein Energy landscape exploration
OPLS	Optimized potential for liquid simulations
CHARMM	Chemistry at Harvard molecular mechanics
AMBER	Assisted Model Building and Energy refinement
SPC	Single point charge
TIP3P	Transferable intermolecular potential 3P
TIP4P	Transferable intermolecular potential 4P
SASA	Solvent accessible surface area
NM	Normal modes
NMA	Normal mode analysis
ANM	Anisotropic network model
TN	Truncated Newton
OBC	Onufriev-Bashford-Case
VDW	Van der Waals
APBS	Adaptive Poisson Boltzmann solver
EC	Exponential contact
PME	Particle mesh Ewald

RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
PC	Principal component
PCA	Principal component analysis
MgP	Mg-protoporphyrin IX
Syn	Cyanobacterium *Synechocystis* PCC 6803
T.el	cyanobacterium *Thermosynechococcus elongates*
PMM2	Phosphomannomutase2
PMM_LEIME	Leismania Mexicana
CPT	Cisplatin
CPT1	Cisplatin mono-aquo
CPT2	Cisplatin di-aquo
MMPBSA	Molecular mechanics Poisson Boltzmann surface area
PMF	Potential mean force
FEP	Free energy perturbation
AFM	Atomic force microscopy
MCPRO	Monte Carlo for proteins
UBQ	Ubiquitin
VB	Virtual bead
Az	Azurin
BOSS	Biochemical and organic simulation system
CRA	Concerted rotations procedure
Fd	Ferredoxin
Fld	Flavodoxin
Hda	Electronic coupling

Chapter 1

Introduction

1.1. Biomolecules: Proteins and DNA

Biomolecules are any molecule present in living organisms, including small molecules like lipids, fatty acids, sterols or vitamins and big molecules such as proteins, polysaccharides or nucleic acids. Nucleic acids and proteins are the most important biological macromolecules. Deoxyribonucleic acid (DNA) is a type of nucleic acid that carries the hereditary material of almost all living organisms while proteins are a type of polypeptides responsible of a vast array of biological functions such as the catalysis of metabolic reactions, molecular transport or DNA replication. Interaction between DNA and proteins during the replication, transcription and translation processes is the central dogma of molecular biology.

Proteins are linear polymers composed by subunits called amino acids. There are 20 different types of standard amino acids and all of them, except proline, have common structural features composed by one alpha carbon, an amino group, a carboxyl group and a variable side chain. Side chains of the amino acids have a great variety of chemical structures and properties. Interactions between side chains are responsible of the three dimensional structure of proteins and its chemical reactivity. Proteins present a wide variety of sizes between a few amino acids to thousands of amino acids. For example, the smallest known protein is Chignolin (Natori 1954, Honda, Yamasaki et al. 2004) with 10 amino acids and the largest one is titin (Natori 1954) with 27000-33000 amino acids; the median length of proteins for Homo Sapiens is 375 amino acids (Brocchieri and Karlin 2005).

Amino acids are linked by chemical covalent bonds called peptide bond. Once an amino acid is linked to another one, we refer to it as a residue, and the alpha carbon, oxygen and nitrogen atoms of the residues are called protein backbone. Due to the double bond nature of the peptide bond the dihedral rotation around this bond is inhibited. Thus, three-dimensional conformations in protein are restricted by this dihedral angle of 180 degrees. The free carboxyl group at the end of a protein is called C-terminus and the free amino group at the beginning is called N-terminus.

In general, proteins fold into unique three-dimensional structures called native conformation. Most of them fold due to the chemical properties of their amino acids but others need molecular chaperons (made also of other proteins) to assist the folding process. Protein structure is decomposed in four levels: primary, secondary, tertiary and quaternary structure. Primary structure corresponds to the linear amino acid sequence without any characteristic shape. Secondary structures are local folds stabilized by hydrogen bonds and repeated along the sequence. The most important secondary structures are alpha helix, beta sheets and turns. Tertiary structure is characterized by interactions between secondary structures in the protein sequence by salt bridges, hydrogen bonds, disulphide bonds and others. This structure controls the basic protein functions. Quaternary structure corresponds to an arrangement composed by two or more proteins in tertiary structure connected by noncovalent interactions forming a protein complex (see Figure 1 right).

Watson and Crick at the university of Cambridge discovered DNA structure in 1953 (Watson and Crick 1953). Two polymer strands coiled around each other forming a double helix DNA, where each strand is a polynucleotide, composed by nucleotide subunits. Each nucleotide is formed by a nitrogenous base (adenine, thymine, guanine and cytosine), a deoxyribose (monosaccharide sugar) and a phosphate group. Nucleobases are linked to form the strand by strong covalent bonds called phosphodiester bonds. Phosphodiester bonds are generated between the phosphor of the phosphate group of one base and the oxygen of the deoxyribose of other base. DNA backbone corresponds to the phosphate group and the deoxyribose of each base.

DNA polymer chains are very stable and (might be) large molecules containing millions of nucleotides where the double helix is stabilized by the hydrogen bonds between base pairs and the base stacking interactions between aromatic bases. The nucleobases can be only paired as adenine-thymine and cytosine-guanine (base pair rule) to make the double stranded DNA. This structure presents two different spaces between the two strands along the double helix called major groove and minor groove. Major groove is 22 Å wide and minor groove is 12 Å wide, and this difference makes more exposed to the solvent and proteins (such as transcriptor factors) the major groove.

DNA can adopt different conformations such as A-DNA, B-DNA and Z-DNA (see Figure 1 left), characterized by small differences in the topological parameters between base pairs (tilt, roll, twist…). Moreover, conformations further depend on environmental conditions such as hydration level, DNA sequence, ions (salt) concentration, etc.; B-DNA is the most common DNA conformation found in cells.

DNA stores the information to code proteins by transcription and translation processes. Moreover, one of the most important DNA function is its replication, which allows cellular division involved in the reproduction of living organisms.

Due to the complexity and functions of proteins and DNA, databases have been created to store and easily access different data from them. The most famous database is the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000), which contains thousands of biological macromolecular structures in continuous expansion.

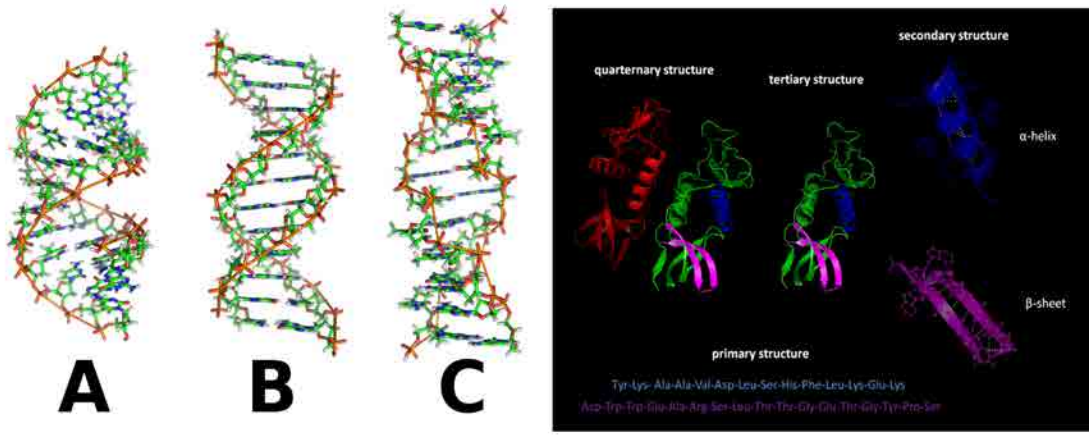


*Figure 1. Left figure corresponds to the canonical DNA conformations A-DNA (A), B-DNA (B) and Z-DNA. Right figure shows the protein structures from primary to quaternary. (**www.wikipedia.org**)*

1.2 Drug interactions with Biomolecules

In biochemistry and biophysics ligands are defined as substances that form a complex with a biomolecule, usually called receptor, modifying its activity. Ligand binding is produced by intermolecular interactions such as ionic bonds, hydrogen bonds, electrostatic and Van Der Waals forces. Ligand binding process is usually reversible (Segal 1975) because these interactions allows dissociation of the ligand. On the contrary, covalent bonding between ligands and receptors produces an irreversible binding (Adam, Cravatt et al. 2001), not so often found in biomolecules.

In the case of protein receptors, ligands are other proteins or small molecules or ions acting as a signal triggering molecule binding in specific regions, the binding site. For DNA, ligand can be an ion, small molecule or a protein binding in a specific region of the double helix.

The strength or tendency of the ligand to bind a protein or DNA fragments is called binding affinity. Reversible ligand binding is an equilibrium process where transition rates are able to characterize ligand affinities. High affinities are produced by strong ligand interactions with the receptor in the binding site (low binding affinity being the opposite), and are related with a long residence time of the ligand in the binding site.

Pharmacology is the science focusing on the interactions between ligands (often called drugs) and biological systems. Protein/DNA ligand interactions play an important role for drug discovery, understanding them is very important for the treatment and management of diseases. For this reason, pharmaceutical companies are interested in the study and improvement of ligands for medical treatments.

1.2.1 Protein-ligand interactions

Protein functions are determined by their tertiary or quaternary structure. These three dimensional conformations, driven mostly by hydrogen bond and hydrophobic interactions, determine to a large extent the protein´s function. Within this structure we typically find a small region, the active site, centering the protein function; active sites often regulate this function through substrate (ligand/cofactor) binding. Ligand binding at the active site can introduce extra interactions modifying the conformational shape of the protein. Designing a drug, then ”reduces” to obtaining a ligand that will replace the natural substrate switching off or changing the protein function.

While most of drugs are designed to directly affect the active site, some proteins present some binding regions called allosteric regions, which are important to regulate ligand binding in the active site but could be significantly far from it. When a ligand binds in the allosteric region a conformational change in the protein activate or deactivate the binding site. Allosteric site is not involved directly in the protein functions but allows controlling the activity of the protein.

As mentioned, drug binding into proteins might be irreversible or reversible. One example of covalently bound ligand is Aspirin molecule (Imanishi, Morita et al. 2011) which binds to a serine residue in the binding site of the cyclooxygenase-2 (COX-2) enzyme. The French chemist Charles Gerhardt discovered aspirin in 1853 and it has been widely used by pharmaceutical companies. Darunavir (Ghosh, Dawson et al. 2007) is an example of ligand bound by strong non bonding interactions to the protease enzyme from several HIV strains.

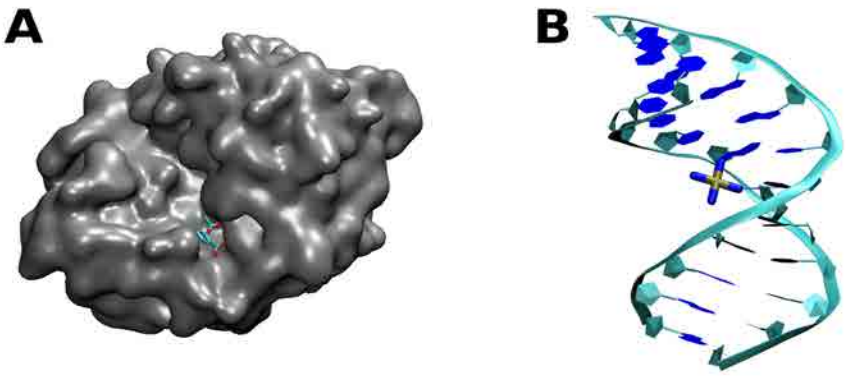


Figure 2. Panel A, crystal structure of Aspirin ligand bound to a protein receptor (PDB ID: 4NSB). Panel B, crystal structure of cisplatin ligand cross-linked to a DNA fragment (PDB ID: 3LPV).

1.2.2 DNA-ligand interactions

There are three main different ways for small molecules to bind double strand DNA: groove binding, intercalation between two base pairs and covalent binding to the bases (Hurley 2002). Besides, some small molecules have more than one way to bind DNA. Groove binding is mostly driven by the hydrogen bonds and ionic interactions between molecules and DNA bases. Intercalation of a small molecule between two adjacent base pairs often requires a planar aromatic system that performs stacking (π-π) interaction with DNA. Moreover, many of these ligands are charged positively due to the importance of coulomb and cation-π interactions in the binding process (Hannon 2007). Moreover, intercalation and the other binding modes are associated to DNA conformational changes (Lerman 1961) modifying the interaction with cell proteins and affecting biological functions. For this reason, intercalator ligands are a potential target for new anticancer drugs. As in proteins, DNA-drug interactions can also be associated to irreversible binding, the three types of covalent binding being: monoalkylation, interstrand cross-linking and intrastrand cross-linking (Goldacre, Loveless et al. 1949). Irreversible binding in DNA is associated to important processes in pharmacology such as apoptosis (cell death).

1.2.3 Experimental methods to study biomolecular interactions

Due to its importance in treatment and management of diseases, understanding the drug binding mechanism is essential. X-ray diffraction and nuclear magnetic resonance (NMR) are two methods focused on the resolution of the crystallographic structures. These methods provides a three dimensional description of atomic positions in biomolecules, allowing to identify the binding site and ligand orientation when applied to protein/DNA-ligand

complexes. They provide useful information about the residues/nucleobases involved in binding interactions and of (possible) conformational changes produced by the binding process if an unbound reference is available.

On the other hand, protein-ligand affinities can not be extracted from crystallographic structures because these techniques are not providing an observable correlated with the ligand affinity. In this way, Surface Plasmon Resonance (SPR) and Isothermal titration calorimetry (ITC) are among the main techniques used to study ligand affinities.

Finally, in this chapter we will introduce atomic force microscopy (AFM) techniques, used in study of protein stretching in chapter 4.

1.2.3.1 X-ray diffraction

X-ray diffraction is the most used technique to determine the three dimensional conformation of large biomolecules such as proteins or DNA with atomic resolution. The first large biomolecule resolved by this technique was the sperm whale myoglobin by Sir John Cowdery Kendrew in late 1950s (Kendrew, Bodo et al. 1958).

X-ray diffraction method is based on three main steps. The first step is obtaining a crystal with the biomolecule to study. Normally, this is the most difficult step and the crystal must be larger than 0.1 mm with a pure composition without imperfections to avoid artifacts in the diffraction pattern. Second, the crystal must be put under an intense monochromatic X-ray beam to produce a diffraction pattern of the reflections. Crystal must be rotated gradually until the original diffraction pattern disappears and all the diffraction patterns corresponding to the different orientations are collected. Last step consists of a computational combination of all the collected patterns to generate and refine the model of the arrangement of atoms in the crystal.

This technique has three main limitations (besides the difficulties in obtaining crystals). First, the size of the molecule (among other things) determines the resolution of the experiment and large molecules tends to produce bad resolutions in the diffraction patterns showing the atoms as tubes of electron density instead of blobs of electron density. Second, hydrogen atoms are not typically resolved in crystals because it is based on the electron density and hydrogen atoms only contain one electron. This aspect might be important, for example, in assigning protonation states in histidines, leading to potential difference in the total charge in the protein. Finally, and possible the most important, the tight and symmetric packing in the

crystal lattice might introduce conformational artifacts due to crystal contacts with neighbours chains. This last effect could significantly change domain positions, loops and side chains (Kopecna, Cabeza de Vaca et al. 2015).

1.2.3.2 Nuclear Magnetic Resonance spectroscopy

Nuclear magnetic resonance (NMR), based on the magnetic properties of the atomic nuclei, was first described and measured by Isidor Rabi in 1939 (Rabi, Millman et al. 1939). It uses the absorption spectra associated to the transition between the nuclear spin levels in the atomic nuclei with an even number of protons or neutron with a strong and external magnetic field (see Figure 3B).

NMR spectroscopy can provide information about structure, dynamics, reaction rate and chemical environment. Samples can be in a solid or liquid state at room temperature but with a reduced temperature results are more precise. As the sample can be in a liquid state, NMR can produce an ensemble of configurations for each one. It is crucial experimental information to study the movements of the protein domains producing a significant advantage respect to X-ray diffraction. In addition, the observed chemical shift produced by the adjacent bonding electrons allow us to identify the hydrogen type information. NMR technique is not precise to resolve large biomolecules structures (or fragments), and it has to be combined with more sophisticated methods.

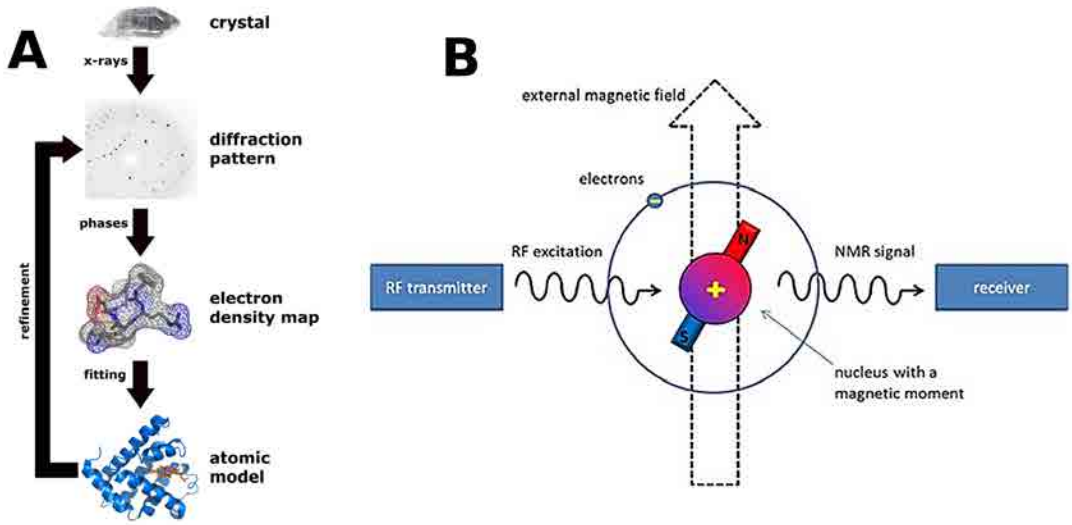


Figure 3. Panel A, X-ray diffraction protocol for the resolution of biomolecules (www.wikipedia.org). Panel B, NMR spectroscopy scheme. (https://www.utu.fi/en/units/sci/units/chemistry/research/mcca/PublishingImages/Applied%20NMR%201w%20480.jpg)

1.2.3.3 Surface Plasmon Resonance

Surface plasmon resonance technique is used to evaluate equilibrium dissociation constant k_d between a receptor and a ligand measuring association and dissociation rates.

The oscillation of conduction electrons between a negative and positive permittivity material stimulated by light is called SPR. When light is emitted and reflected in the interface of metal/dielectric or metal/vacuum parallel electromagnetic waves are generated called surface plasmon polaritons. Oscillations are very sensitive to the change in this boundary because the movement of these waves is located in the surface. When a surface plasmon interacts with an irregularity or a particle the energy is re-emitted as light that can be detected (Zeng, Baillargeat et al. 2014) (Scheme in Figure 4A).

This technique has been applied to measure association/dissociation rates of binding processes determining the resonance difference produced by the unbound and bound monomers.

1.2.3.4 Isothermal titration calorimetry

Isothermal titration calorimetry is used to determine thermodynamic parameters of interactions in solutions for small and large molecules such as drugs or proteins. It is a quantitative technique able to determine binding affinities and enthalpy changes of the interaction between one or more molecules (Pierce, Raman et al. 1999).

Isothermal titration calorimeter consists of two identical cells surrounded by an adiabatic jacket where a constant power is applied to keep the same temperature in both cell. One cell is used as the reference containing a buffer or water and the other is filled with the macromolecules to study. Ligand is added with small controlled quantities and depending of the reaction nature (exothermic or endothermic) the power must be changed to reach the same temperature in both cells. Power difference measured (heat difference) to maintain the same temperature along each ligand injection allows us to obtain the heat exchange due to the macromolecule ligand binding process (see scheme in Figure 4B).

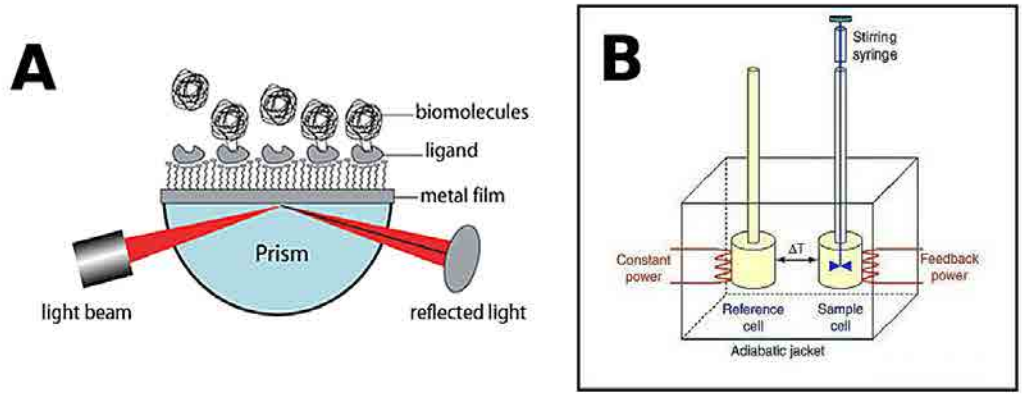


Figure 4. Panel A, surface Plasmon Resonance (SPR) experimental method scheme. (http://www.biosensingasia.com/images/ how_does_spr_work_figA.png). Panel B, isothermal Titration Calorimetry (ITC) scheme. (http://pharmaxchange.info/press/wpcontent/uploads/2012/08/isothermal_calorimetry_itc_instrumentation.png)

1.2.3.5 Atomic Force Microscopy

Atomic Force Microscopy (AFM) technique was developed at IBM Research - Zurich by G. Binning in 1986 (Binnig, Quate et al. 1986). It was originally developed to measure a roughness of a surface at a high resolution but nowadays has been applied to unfold proteins and measure the molecular forces along the process. As mentioned before, three dimensional protein structures are related to the protein functions and a huge number of diseases such as Parkinson or Alzheimer are related with protein misfolding; studying the energy landscape of the folding pathway might help to understand the process that can lead to these diseases.

AFM unfolding experiments uses a sample composed of proteins (in a solution) and attached to a surface. During the experiment, the microscopic cantilever arm is repeatedly introduced into the solution. During these cycles, the cantilever tip can be attached randomly to a molecule of the solution and the molecule is stretched by pulling out the cantilever. Protein's reaction force can be measured using the bending produced in the cantilever during the pulling process (See Figure 5).

AFM has been used to measure the force-extension curves for a wide set of proteins such as titin constructs (Li, Linke et al. 2002, Linke, Kulke et al. 2002, Bullard, Ferguson et al. 2004), ubiquitin (Carrion-Vazquez, Li et al. 2003), and azurin unfolding (Giannotti, Cabeza de Vaca et al. 2015).

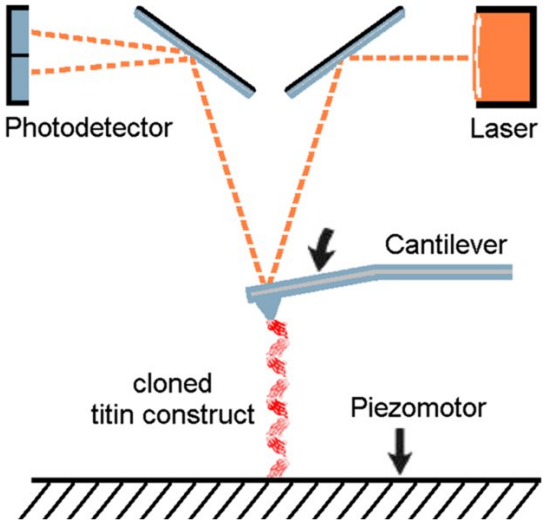


Figure 5. Atomic Force Microscopy device. (http://www.uni-muenster.de/Biologie.AllgmZoo/AG-Linke/images/AFM_1.JPG)

1.2.4 Disadvantage of experimental techniques

Experiments are able to provide very important information about thermodynamics properties such as affinities and structural information, crystal structures, of molecular systems. Are they enough to understand protein/DNA ligand recognition mechanisms? How easy is to get such information?

There are thousands of degrees of freedom in protein/DNA ligand interactions. Experimental techniques are based on data fitting and interpretation. Moreover, obtaining such information is time consuming and often associated to a high cost. In addition, an atomic descriptions of the system dynamics is difficult; experiments provide important information for drug development but nowadays still no experimental technique is able to provide robust information about the time evolution (dynamics) of the interaction mechanism at the atomistic level of description.

In this way, methods to study protein/DNA ligand interaction mechanisms at atomistic detail are extremely valuable for drug design targeting proteins or DNA. During the last years, computational methods based on molecular mechanics or quantum mechanics have been applied to connect experimental information with all atom descriptions of the phenomenon.

1.3 Theoretical methods

Atoms connected by covalent bonds determine the molecular chemical structure. These atoms are not in relative fixed positions due to the thermal energy (and to less degree to the quantum zero point energy). It produces relative movements between molecules and vibrational movements of bonds, angles and dihedrals of the molecule. Thus, molecules are dynamic systems with thousands of possible configurations where some configurations are more favorable than others due to the free energy associated to the conformation.

Physics provides two well defined approaches to study molecular systems in standard conditions: classical and quantum mechanics (QM). Selection of the model depends of the phenomena to study and the size of the system. QM provides accurate atomic interaction descriptions taking into account the electronic distribution but resolution of these equations makes this approach expensive for large systems (more than 100/1000 atoms). On the other hand, classical mechanics offers a good approach to study large molecular systems but it neglects quantum effects associated to the molecular interactions (electronic effects). A brief description of these methods will be provided in the following sections.

Classical mechanics (also known as Newtonian mechanics) is used to describe the movement of macroscopic objects. Molecular mechanics (MM) uses classical mechanics for the atom dynamic description using classical force fields. At MM level, atoms are considered charged spheres connected by spring bonds. Force fields connect spatial atomic positions in a molecule with an energy, defining a Hamiltonian based on bond distances, angles, dihedrals, Van der Waals and electrostatic interactions (Lewars).

QM solves the Schrödinger equation to get the energy, the wave function (the distribution of the electrons) and its derivative, obtaining gradients of the atoms in molecular systems to model molecular motion. Resolution of Schrödinger equation for molecular systems is very expensive for small molecules since there is not exact solution for systems with more than one electron; several approximations have been developed to solve the Schrödinger equation for these systems. For small molecules, detailed QM methods can take days and quick approximations just minutes or hours. In any case, even the fastest methods are not applicable to study the dynamics of proteins or large DNA fragments due to the computational cost associated to each calculation.

1.3.1 Multiscale approaches

Multiscale modelling tries to solve problems with different features at different time/space scales. In physics and chemistry, multiscale modelling aims to extract material properties or system behavior using models focus on each information level. Four theory levels are mainly distinguished in the study of molecular systems by the use of time/space criteria. First level are quantum mechanics models where the electronic information of the atoms is included to study phenomenon such as chemical reactions or electron transfer processes. Second level corresponds to molecular mechanics models where information about atomic positions is the base of the models. These models are used to study at atomistic detail the conformational changes of biomolecules or protein/DNA-ligand dynamics at nanosecond or microsecond time scale. Hybrid methods combining MM with QM have been successfully applied to the study of macromolecular systems (Warshel and Levitt 1976). Third level, called coarse-grained level, includes information about atomic groups to reduce the number of degrees of freedom associated to the molecular dynamic level. The range of the coarse-grained models is quite diverse, going from few atoms, ~1nm, to 100 nm (typical virus) or 1000 nm (bacterium size). One application of these models is the study of nano particles or interactions dynamics between large protein clusters. Fourth level is the mesoscale/continuum model where the system is studied as a continuum mass instead of a group of particles. These models have been applied to study fluid dynamics of different materials.

This thesis is mostly focused on the study of molecules at the atomic level from a computational point of view. In the last chapter, we have studied a coarse grained model to accelerate the sampling process in protein-protein docking problems. In the below section, detailed explanation about methods applied in this thesis will be expanded.

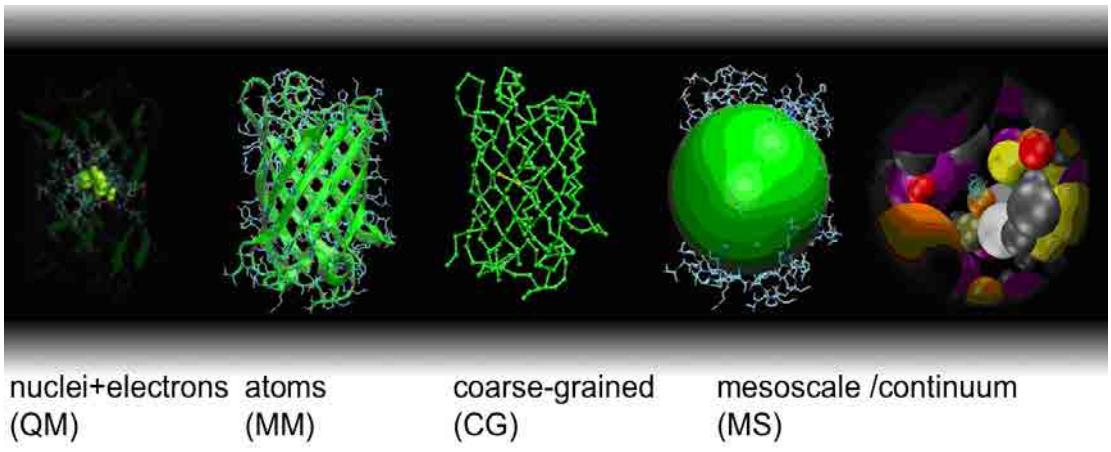


Figure 6. Representation of the different scales in the multiscale approaches for molecular systems. (http://www.nano.cnr.it/upload/allegati/allegato /247.png)

1.3.2 Coarse grained models

In molecular simulations, coarse grained (CG) methods are based on the replacement of a group of atoms by a larger particle averaging or smoothing the properties of these atoms. On a general rule, the larger is the number of atoms being replaced by a single particle, the more approximate (less detailed) is the CG model. More approximation obviously goes together with less degrees of freedom (compared to an equivalent all-atom model) and faster conformational sampling. In this way, CG models have been developed to study biological processes that require very large systems or longer time scales. There are many ways to reduce the degrees of freedom in a CG model. Depending of the system to study and due to the elimination of fine interaction details different models have been proposed.

Residue based models represent each residue as one or more CG beads trying to keep the total residue properties (identity). Scorpion CG model (Basdevant, Borgis et al. 2012) is an example of residue based model focus on protein association processes. One of the most popular CG models, the MARTINI force field by Marrink (Marrink, De Vries et al. 2004), was first proposed to study lipids. An extension for lipoproteins of Marrink's CG model has been successfully applied to describe structural properties and formation of high-density lipoprotein particles (Shih, Arkhipov et al. 2006) and the assembly of proteins and lipids in lipoproteins particles (Shih, Arkhipov et al. 2007).

Shape based coarse graining models have been developed to study large conformational changes in huge molecular assemblies. This CG model is based on the molecule shape to generate CG beads and extract properties trying to keep an accurate description of the geometry for the vast shape variety in molecules (compact domains, tails...). One example of successful application of a shape based CG model is the description of the stability and dynamics of a virus capsid (Arkhipov, Freddolino et al. 2006) or the rotating bacterial flagellum by molecular dynamics (Arkhipov, Freddolino et al. 2006).

Some processes like protein/DNA ligand binding require an atomistic representation to reproduce system dynamics. Coarse-grained models are not able to reproduce hydrogen bond interactions or protonation states. In these cases, hybrid methods CG/all atom are applied to perform long simulations of the whole system in CG representation but keeping a detailed all atom representation of the important regions.

1.3.3 Force Fields

In molecular mechanics, force fields correspond to an energy function and a set of parameters used to compute the potential energy (and its derivatives) of a molecular system at all atom or coarse grained level. Typical force field parameters for each atom types are mass, van der Waals radii and charges. For covalently bonded atoms, the force field provides structural parameters derived from experimental measures (empirical) or quantum mechanics evaluations.

The energy terms in molecular mechanics are split in two parts: bonded and non bonded interactions. Bonded interactions are terms associated to the atoms that are linked by covalent bonds. Non bonded interactions are terms associated to long range interactions. Due to the scalar nature of the energy, total energy of a system corresponds to the sum over all the interacting terms.

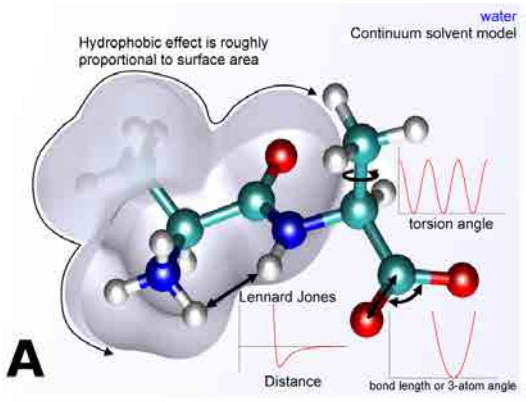
Normally, covalent interactions are described by three terms: bond, angle and dihedral term. Bond term is described with a harmonic potential (quadratic term) of the difference between the ideal interatomic distance and the distance. This model is missing bond-breaking possibility but works very well close to the equilibrium distances. Angle term is also model by the same harmonic potential based on the equilibrium angle. Dihedral torsional parameters are more variable because the function must be periodic and able to provide more than one minima for the different torsional angles. Additional improper dihedrals are added in force fields to enforce planarity in some structures such as aromatic rings or other conjugate systems.

Nonbonding terms describe long-range interactions between atoms. These terms are usually described by pairwise energies and, due to the huge number of possible pairs, they are the most expensive part in the energy evaluation. Typically, the non bonding part is defined as a sum of two terms: Van der Waals and electrostatic term. Van der Waals interaction normally is modeled by a Lennard-Jones potential and electrostatic term is modeled by the Coulomb potential. Addition of a polarization term in force fields, to take into account charge redistribution due to the position of the other system charges, is usually neglected because it high computational cost. Nevertheless polarizable force fields, such as the AMOEBA one, have reported higher accuracy in modelling (Ponder, Wu et al. 2010).

Popular force fields developed specifically for macromolecules and used in molecular simulations are AMBER (Assisted Model Building and Energy refinement) (Cornell, Cieplak

et al. 1995), OPLS (Optimized Potential for Liquid Simulations) (Jorgensen and Tirado-Rives 1988) and CHARMM (Chemistry at Harvard Molecular Mechanics) (Karplus 1983). Since the firsts versions, different updates have been released adding improvements in the parameters for better energy evaluations.

In particular, AMBER parm99sbBSC0 force field (Pérez, Marchán et al. 2007, Ivani 2015) was specifically developed to improve DNA parameterization adding a new atom type in DNA backbone to correct two torsional terms. This new version has demonstrated a great stability for long computational simulations.



$$
\begin{aligned}
U &= \sum_{\text{bonds}} k_r (r - r_0)^2 \\
&+ \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\
&+ \sum_{\text{dihedrals}} k_\phi \left[1 + \cos(n\phi + \phi_0)\right] \\
&+ \sum_{\text{atom } i} \sum_{j \neq i} 4\epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \\
&+ \sum_{i} \sum_{j \neq i} \frac{q_i q_j}{\epsilon_0 r_{i,j}}
\end{aligned}
$$

B

Figure 7. Panel A, diagram of the molecular mechanics potential terms. (https://upload.wikimedia.org/wikipedia/commons /5/5c/MM_PEF.png). Panel B, molecular mechanics potential energy equations for AMBER force field decomposed by potential terms. (http://www.chem.hope.edu/~krieg/shorb/amberff.png)

1.3.4 Solvent models

Water molecules and ions dissolved in water are very important for biomolecular studies in living organism where all chemical reactions are produced in a water medium. For macromolecules, like proteins or DNA, solvent properties like pH or temperature determine the three-dimensional conformation or surface residues protonation. These properties are crucial for protein and DNA functions in cells. For this reason, inclusion of solvent interactions in molecular mechanics is mandatory to perform accurate simulations.

Computational water models have been developed in order to model the solvent effects, often focused on reproducing specific properties such as the heat capacity. In general, these models can be split in two groups: explicit and implicit solvents. Explicit solvents use thousands of solvent molecules (typically waters and ions) and it is the most frequent solvent model in molecular simulations. Due to the huge number of solvent molecules needed in these

simulations the convergence is slow and the CPU time for treating the solvent overcomes the solute one. For this reason, another solvent approximation less computationally expensive, the implicit solvent, has been developed. Implicit solvent approximations treat the solvent as a continuous medium, including an additional force field term describing average properties of the real solvents, and considerable reducing the degrees of freedom and computational time.

The most frequent water models for molecular mechanics simulations in explicit solvents are SPC (Berendsen 1987), TIP3P and TIP4P (Jorgensen, Chandrasekhar et al. 1983) due to the high balance between performance and accuracy of the results. These models corresponds to the so called three site models were each atom has a partial charge and the oxygen atom has the Lennard-Jones parameters. Difference between three sites water models comes from the oxygen-hydrogen distance, hydrogen-oxygen-hydrogen angle, van der Waals parameters and charge distribution.

Implicit models are based on approximations of the Poisson-Boltzmann (PB) equation (Baker 2005). PB equation is a second order non-linear partial equation (see equation 1) and describes the distribution of the electrostatic potential in solution in the normal direction to a charged surface in the presence of ions. As PB equation is expensive from a computational point of view, Generalized Born (GB) approximations (Still, Tempczyk et al. 1990) have become popular in the recent years in molecular dynamics (MD) applications (Dominy and Brooks 1999, Calimet, Schaefer et al. 2001, Gallicchio and Levy 2004). GB method is an approximation to the exact linearized PB equation and it is based on model the solute as spheres with a fixed dielectric constant. For molecules, each atom has an equivalent sphere and the radius of the spheres is called Born radius or alpha radius. PB equation is:

$$\nabla[\epsilon(\mathrm{r})\nabla\phi(r)] = 4\pi\rho(r) + \kappa^2\epsilon(r)\phi(r) \quad (1)$$

Where ϕ is the electrostatic potential, ϵ is the position dielectric constant, κ is the Debye-Huckel parameter $\kappa \sim \sqrt{[ions]}$ and $\rho(r)$ is the molecular charge distribution.

The most used GB model is called generalized born surface area (GBSA) (Qiu, Shenkin et al. 1997) where the free energy of transferring the molecule from vacuum into solvent (solvation free energy ΔG_{solv}) is modeled by two terms: $\Delta G_{nonpolar}$ and ΔG_{el} (Onufriev 2008). $\Delta G_{nonpolar}$ corresponds to the free energy to solvate the molecule with no charges and it is composed by ΔG_{CAV} of solvent-solvent and by the ΔG_{VDW} of solvent-solute Van der Waals interactions. ΔG_{el} corresponds to the solute-solvent electrostatic interaction (see

equation 2). Nonpolar terms are modeled proportional to the solvent accessible surface area (SASA) and are empirically determined, and ΔG_{el} is computed using the GB solution. For this last term, several approximations have been derived.

$$\Delta G_{solv} = \Delta G_{nonpolar} + \Delta G_{el}$$
$$\Delta G_{nonpolar} = \Delta G_{CAV} + \Delta G_{VDW} \quad (2)$$

GB models evaluate the electrostatic part of the solvation free energy as a pairwise sum between atomic charges. These pairwise interactions can be modeled with the formula introduced by Still (Still, Tempczyk et al. 1990) for molecules with a constant dielectric of 1.

$$\Delta G_{el} \approx -\frac{1}{2}\left(1 - \frac{1}{\varepsilon_w}\right)\sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j e^{-\left(r_{ij}^2 / 4R_i R_j\right)}}} \quad (3)$$

Where r_{ij} is the distance between atom i and j, $q_i q_j$ are the partial charges and $\varepsilon_w \gg 1$ is the dielectric constant of the solvent. R_i and R_j are the effective born radii of the atoms which depends of the intrinsic atomic radii and the relative position of the other atoms in the system. Effective Born radii can be estimated by different approaches such as coulomb field approximations (Scarsi, Apostolakis et al. 1997, Ghosh, Rapp et al. 1998) or continuum dielectric models (Schaefer and Karplus 1996). Approximations based on the fitting of adjustable parameters are often added to these methods to speed up the Born radii evaluation (Hawkins, Cramer et al. 1996, Onufriev, Bashford et al. 2004). These methods are less general than the standard formalisms but make faster simulations of proteins and DNA molecules.

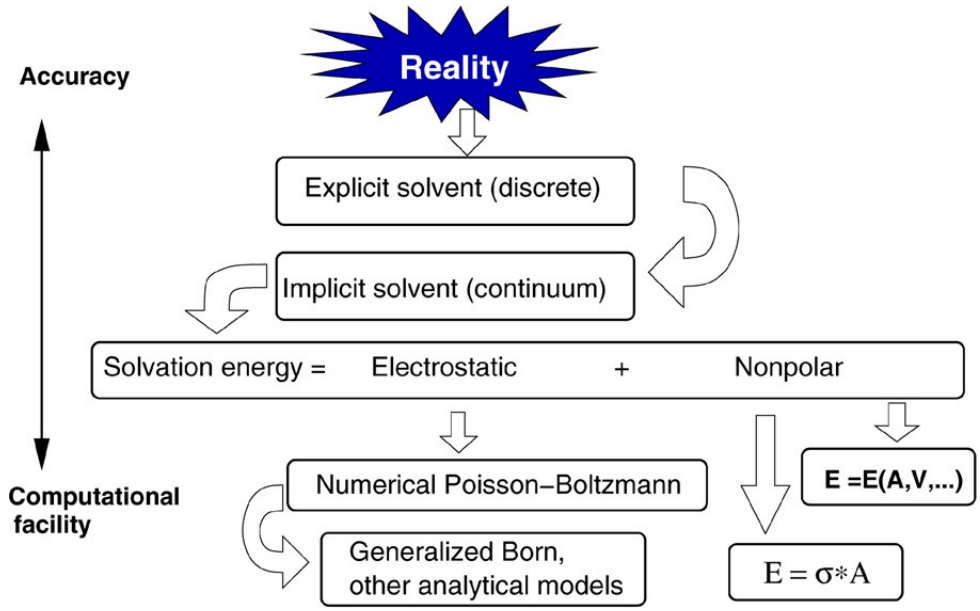


Figure 8. Implicit solvent framework (Onufriev 2008).

1.4 Computational methods to study Protein, DNA and ligand interactions

Some methodologies have been developed using these theoretical models to describe the interactions in biomolecules and between biomolecules and ligands. A main area of application for molecular mechanics is the optimization of structures. Force field and solvent models are combined to find a close conformational local minima using algorithms such as steepest descend (Arfken 1985) or Truncated Newton (Nash and Nocedal 1991). Global minima search requires more sophisticated algorithms like simulated annealing (Kirkpatrick, Gelatt et al. 1983), metropolis criterion (Metropolis, Rosenbluth et al. 1953) or other Monte Carlo (MC) methods.

One of the most important applications of force fields and solvents are MD and MC methods. MD is based on the integration of the Newton equations, and it allows us to simulate with atomistic detail the dynamics of molecular systems. On the contrary, MC methods use random movements to sample the conformational space. Both methods are described in detail in the sections 1.4.1 and 1.4.2.

Molecular mechanics energy functions combined with MD and MC algorithms have been used to calculate, for example, binding constants (Kollman, Massova et al. 2000, Huo, Massova et al. 2002), protein folding kinetics (Snow, Nguyen et al. 2002), protonation

equilibrium (Barth, Alber et al. 2007), active site coordinates (Mobley, Graves et al. 2007) and design binding sites (Boas and Harbury 2008).

1.4.1 Molecular Dynamics

Molecular dynamics is a computer simulation technique based on the physical movements of atoms and molecules following Newton's equations. This technique simulates time dependent processes in the range of the time scale available by the computational resources. Indeed, MD is able to reproduce from femtoseconds /picoseconds phenomenon corresponding to local motions like atomic fluctuations to milliseconds simulating the folding/unfolding of small molecules (Lindorff-Larsen, Piana et al. 2011).

In MD, Newton's second law (force = mass · acceleration) is used to simulate the dynamics of each atom in molecular systems with the general formalism described by the equation 4:

$$F = -\nabla \mathrm{V}(\mathrm{q}) \tag{4}$$

Where the force F in a particle is related with the gradient of the potential (V). At the beginning of an MD simulation, a random distribution of velocities is assigned to each atom with an average velocity determined by the temperature according to the equipartition theorem. Then, the forces that each atom exerts on each other determine module and direction of the velocity during the time system evolution. MD simulations have to be equilibrated at the beginning to avoid numerical instability due to the initial coordinates (typically from a high energy potential state). In general, this process is based on an initial minimization and a heating process where temperature starts from 0 and goes to the desired temperature (Normally 300 K). Typically, in heating processes a few picoseconds are enough at each constant temperature to equilibrate the system and jump to the next temperature avoiding instabilities. In addition, integration time step used in atomic detailed MD is kept around 1-4 femtoseconds, to avoid numerical integration instabilities.

Most of the MD softwares use molecular mechanics force fields as the potential to compute forces between atoms in the simulation. Each integration of the movement with the time step equation describes the position and velocities of the particles and the group of snapshots at different times is called trajectory. In practice, Newton's equations are not used to integrate the movement due to numerical instabilities. Some approximations have been developed to avoid these problems such as Verlet integration (Verlet 1967) or velocity Verlet (Swope, Andersen et al. 1982). In general, MD simulations use explicit solvent models defining boxes

where periodic boundary conditions are applicable such as cubic or ortorombic boxes and the size depends of the system. Also, implicit solvent models can be used for speed up MD simulations.

The most famous MD softwares are NAMD (Phillips, Braun et al. 2005), GROMACS (Hess, Kutzner et al. 2008), Desmond (Bowers, Chow et al. 2006) and AMBER (Case, Cheatham et al. 2005). They contain optimized codes able to take the maximum advantage of the computational resources; have been implemented to work in parallel with computer clusters and accelerated devices such as GPU clusters. Currently, MD codes are able to simulate up to ~1-10 microseconds on current supercomputers (Freddolino, Liu et al. 2008). Using a special purpose machines, the D.E Shaw research lab has been able to push this limit into hundreds of microseconds (Freddolino, Liu et al. 2008, Götz, Williamson et al. 2012) and it has been able to simulate the folding of a WW domain (Shaw, Maragakis et al. 2010) (Shaw, Dror et al. 2009).

Accelerated Methods

In molecular simulations, MD and MC methods have been developed to study a large part of the conformational space. The huge amount of local minima in the energy landscape produces problems to jump between minima in a feasible computational time. For this reason, some methods have been developed aiming to accelerate the landscape exploration of the local minima. The most famous are Metadynamics (Laio and Parrinello 2002), Umbrella Sampling (Torrie and Valleau 1977), Replica exchange (Earl and Deem 2005) and Steered molecular dynamics (Israelewitz, Gao et al. 2001).

Metadynamics technique uses a set of variables to describe the system called collective variables. In each simulation step, a gaussian function is added to the energy landscape for each collective variable explored. During the simulation, the sum of gaussian functions reduces the probability to comeback to the same position biasing the simulation to explore another minima. When collective variables start to fluctuate heavily, the free energy generated with gaussians becomes constant and it means that the energy landscape can be reconstructed as an inversion of the sum of gaussian functions. Umbrella sampling technique uses the inclusion of a biasing potential in a reaction coordinate to explore the conformational space along the reaction coordinate. The free energy profile of the system can be reconstructed directly extracting the contribution of the biasing potential. Replica exchange method (also called parallel tempering) consists of the simulation of N different system replicas, each with a different simulation parameter (typically temperatures, but also

Hamiltonian, pressure, etc.). Then, at some point, replicas exchange the configurations using a metropolis criterion; statistics are collected for the reference replica (unperturbed Hamiltonian, room temperature, etc.). This technique allows a better exploration of the conformational space. Steered molecular dynamics technique applies forces to particles of the system in a desire direction of the movement to measure the system response. These forces are generated using harmonic constraints attached to virtual beads located and moved in a specific point and direction of the space. Two main strategies for the bead perturbation have been developed for steered molecular dynamics: constant velocity and constant force. For instance, steered molecular dynamics with constant velocity has been applied to study the unfolding force of protein systems (Carrion-Vazquez, Li et al. 2003).

1.4.2 Monte Carlo Methods

Monte Carlo (MC) methods are based on random sampling of events to obtain statistical results. In theory, MC methods have a highest convergence than deterministic methods to achieve the results (Newman and Barkema 1999) when the number of degrees of freedom is large. In practice, however, we find limited MC simulations in complex biomolecules: most studies are performed applying MD techniques. MC methods have been applied in different scientific areas including astrophysics, aerodynamics, fluid dynamics, telecommunications, and optimization problems. For example, these methods have been successfully applied to solve differential equations (Graham, Kurtz et al. 1996), integrals (Hammersley 1960), scattered radiation distribution (Seibert and Boone 1988) and many others. The main part of the MC methods is the generation of an accurate random sequence of values in an affordable time.

MC sampling methods for molecular systems are less computationally expensive than MD because don’t need to integrate the movement equations (gradient calculations, etc.) (Jorgensen and Tirado-Rives 1996). On the contrary, MC methods need to find a set of parameters for the random components able to sample the whole conformational space. The other main problem is the lack of time dependent information of the simulations limiting the results.

The most famous MC algorithm used in molecular simulations is Metropolis MC (Binder and Heermann 2010). Metropolis MC algorithm is used to generate a Boltzmann distribution of the accepted conformations using an acceptance criteria based on the total energy difference between the initial and the final conformation. If energy of the system decrease the step is always accepted but if energy system increases the step is accepted with a probability

proportional $e^{\Delta E/K_B T}$ where ΔE is the energy difference, K_B is the Boltzmann constant and T is the temperature. Typically, acceptance ratio for MC sampling should be around 30%.

The difficulties in designing MC moves in biomolecules, has limited applying these techniques. A collective random move will vanish acceptance ratio. On the other site, local motions, with higher acceptance, will hinder exploration of the whole conformational space. When comparing to MD simulations, we find studies involving MC limited to local sampling. We can underline studies by the Jorgensen group using MCpro (Jorgensen and Tirado–Rives 2005) and by the Essex lab (Michel, Taylor et al. 2006).

Our in house PELE methodology, represents an attempt to circumvent these limitations by adding protein structure prediction techniques into the MC moves (see section 1.5)

1.4.3 Docking methods

Docking methods aim at predicting (discriminating) the correct orientation between two molecules to form a stable complex. Typically the larger molecule is called receptor and the small one is called ligand. This general term is used in complexes with any two biomolecules, although it usually refers to protein/DNA interacting with a small molecule, where each interaction is significant due to the small number of intermolecular contacts.

In general, docking methods are divided in two groups: rigid and flexible docking. Rigid docking uses a fix receptor and ligand conformations, while flexible docking adds some degree of flexibility in the ligand and/or the receptor (mainly just in the binding pocket). The importance of taking flexibility into account depend whether conformational changes during the binding process are more or less critical. Obviously rigid docking is significantly less expensive due to the drastic reduction of the degrees of freedom in the problem, being applied usually to libraries up to few millions of compounds.

The common docking algorithm procedure consists of two main steps: sampling and scoring. Along the sampling step different algorithms generate (hundred of thousands of) ligand poses around the receptor in different orientations. Then, a quick geometric criteria based on grid discretization of atom positions combined with a cross correlation analysis is applied to reduce the number of acceptable poses keeping a few thousands poses. Finally, the scoring step ranks poses using scoring functions developed to take into account the possible interactions with a quick evaluation function. For flexible docking, the prohibited number of

possible conformations requires an intelligent method to select a subset of relevant conformations for consideration.

Due to its importance in pharmacology industry, there are many different docking methods. For protein ligand docking GLIDE (Friesner, Banks et al. 2004, Friesner, Murphy et al. 2006), Swissdock (Grosdidier, Zoete et al. 2011), AUTODOCK (Morris, Huey et al. 2009) and GOLD (Verdonk, Cole et al. 2003) are among the more used ones. Protein-protein docking algorithms like pydock (Cheng, Blundell et al. 2007), Fiberdock (Mashiach, Nussinov et al. 2010), Zdock (Pierce, Hourai et al. 2011) and HADDOCK (Dominguez, Boelens et al. 2003) are also widely used.

1.4.3.1 Scoring Functions

Scoring functions are quick and approximate mathematical function to predict binding affinities between docked molecules. While they mostly refer to small drug compounds and macromolecules like protein or DNA, scoring functions to study protein-protein or protein-DNA affinities have also been developed.

In general, there are three types of scoring functions: force field, empirical and knowledge-based. Force field based scoring functions use a combination of energetic terms (similar to the above introduced molecular mechanics force fields) to evaluate the affinity. Normally, Van der Waals, electrostatic potential, strain and solvation term are included. Empirical scoring functions count the number of interactions between receptor and ligand and the type of interactions (hydrophobic, hydrogen bonds...). Coefficients of these functions normally come from multi linear regressions fits of databases. Knowledge-based scoring comes from statistical observations of large data sets assuming that close intermolecular interactions are related with binding affinity.

Due to the importance for drug design, research focus on the development of efficient protein-ligand scoring functions still being active. Recently, a machine learning algorithm based on random forest algorithms called RF-score (Ballester and Mitchell 2010) have been developed to score protein-ligand binding affinities showing better performance that most of the main scoring functions.

1.5 Methodology: Protein Energy Landscape Exploration (PELE)

As mentioned, application of MC techniques in biomolecule sampling is not a straightforward procedure due to the difficulties in applying global sampling moves. Moreover, the protein-ligand energy landscape is full of local minimums due to the huge number of degrees of freedom. In general, ligands contain rotatable bonds and during the interactions with other molecules they are able to adopt different conformations. On the other hand, protein backbones and sidechains are flexible and introduce a large number of different states. For all these reasons, even with the binding site identified previously, the number of possible configurations to take into account is enormous.

Protein Energy Landscape Exploration (PELE) was designed to enhance MC sampling in such difficult cases. Its key contribution, resides in using protein structure prediction techniques coupled to random trials, enhancing the MC moves toward important sampling regions. Originally developed to sample and study the conformational space between proteins and ligands (Borrelli, Vitalis et al. 2005), PELE has been applied to study numerous ligand migration, induced fit docking and protein dynamics with less computational cost than MD (Borrelli, Cossins et al. 2010).

1.5.1 PELE Scheme

PELE is an iterative MC algorithm based on a combination of structure prediction algorithms, capable of producing rapid and accurate protein or protein ligand conformations. PELE uses the OPLS (Jorgensen, Maxwell et al. 1996) force field to evaluate the total energy of the systems. Solvent contribution is estimated using two implicit GBSA solvent model called SGBNP (Ghosh, Rapp et al. 1998, Gallicchio, Zhang et al. 2002) and VDGBNP (Zhu, Shirts et al. 2007) where a Debye-Huckel term (Edinger, Cortis et al. 1997) have been included to take into account the ionic strength contribution. PELE uses the multiscale non bonding algorithm (Zhu, Shirts et al. 2007) to speed up pair list generation used in the non bonding energy terms. Moreover, cell list optimization is implemented for the nonbonding list updating.

Two main parts compose each PELE MC step: perturbation and relaxation. Perturbation is split in ligand and protein parts. These two perturbations take out system's initial conformation from a local minimum generating another configuration. Then, a relaxation part, including a side chain prediction algorithm and a global minimization, is performed to find another local minimum close to the new perturbed conformation generated. After these

two steps, the final configuration is accepted or rejected using a metropolis criterion (Metropolis, Rosenbluth et al. 1953) to approximate a Boltzmann distribution of the PELE steps in terms of energy. A typical PELE step takes around 1-2 minutes in a standard CPU depending of the system size (PELE scheme is shown in Figure 9).

A more detailed description of the task performed in each step is given here.

1) Perturbation

Ligand Perturbation. This perturbation is performed if a ligand exists (and we aim at sampling its dynamics). Initially, the ligand is perturbed with random rotations and translations. These perturbations can be uniform from 0 to a certain maximum value or uniformly distributed around some value. Selected translation direction can be maintained during more than one step to force exploration over a selected direction. Once the program has perturbed the ligand, it checks for possible steric clashes: i) clashes with backbone atoms discard the ligand's move; ii) clashes with side chains introduce an additional sampling in side chain rotamers (within a chosen distance from the ligand's heavy atoms) and pre-computed rotatable bonds of the ligand itself. The entire process, ligand move plus steric clash relief, is repeated for several trials (user defined) and the lowest energy one is chosen. Ligand movement can be restricted to space regions called boxes to reduce conformational search space. Boxes can be cubic, prismatic or spherical shaped and box size depends of the system.

Protein perturbation. This task aims to induce protein global motion by introducing a backbone perturbation following a displacement along one (or a combination) of the lowest normal modes (NM). Normal mode analysis (NMA) approximates the harmonic nature of global fluctuations through the second derivatives and frequencies. NMA has been applied to molecular force fields and more approximated methods describing complex conformational transitions (Xu, Tobi et al. 2003). For a detailed explanation see chapter 2. PELE uses an NMA version called Anisotropic Network Modeling (ANM) (Doruker, Atilgan et al. 2000), an elastic network model based on a simplified potential connecting neighbor alpha carbon atoms (within a defined distance cut off). The most significant advance is its increase in speed due to the reduction in degrees of freedom reducing drastically the computational time. Another main advantage of this approximation is the no need for structure minimization because the harmonic potential is selected assuming that the distances are in a minimum. Once the perturbation direction has been chosen, PELE applies it, through an all-atom

minimization including a harmonic constraint in the alpha carbons pointing in the NM direction.

2) Relaxation

Side chain sampling. This task has been developed to rearrange (relax) protein side chains close to the ligand (in response to the ligand move) or whose energy has increased considerably along the protein perturbation. Side chain prediction algorithm uses an optimized method to generate non clashing side chain configurations based on stored rotamer libraries (Lasters, De Maeyer et al. 1995, Jacobson, Pincus et al. 2004). The algorithm uses a hierarchical approach of steric filtering, clustering and energy scoring of each cluster representative. This heuristic approach cannot guarantee the global minima exploration (prohibitive in terms of computational efficiency), but provides very good estimates for moderate number of side chains, < 25 residues. For this reason, this iterative procedure is focused in a quick search of a local region.

Global Minimization. Global minimization includes all the atoms involved in the previous steps to perform a minimization with a truncated Newton algorithm. A small harmonic position constraint is (optionally) added in ANM node atoms driving the minimization to keep the perturbed backbone conformation. Moreover, another harmonic position constraint can be added in the ligand atoms to force the protein atoms adaptation to the ligand position. It is useful in cases where binding site is narrow and has to be opened to allow ligand entrance or exit.

1.5.2 Parallel PELE implementation

PELE has been parallelized to speed up a search using a spawning criterion in a collective variable where all independent trajectories, running in different cores, share information. The spawning criterion over a reaction coordinate, such as distance between atoms or RMSD, defines a maximum or minimum value allowed. Then, after each step, PELE checks the value of the reaction coordinate for each independent simulation. If a conformation generated goes out of the spawning criteria range, the coordinates are replaced by the best conformation found (in terms of spawning criteria) using the MPI communication protocol. In Figure 10, for example, we see two processors who are falling behind the “reaction coordinate” criteria, and the leading processor transfer (spawns) its coordinates into them; in the following MC step, these two processors will start with the best coordinates ones.

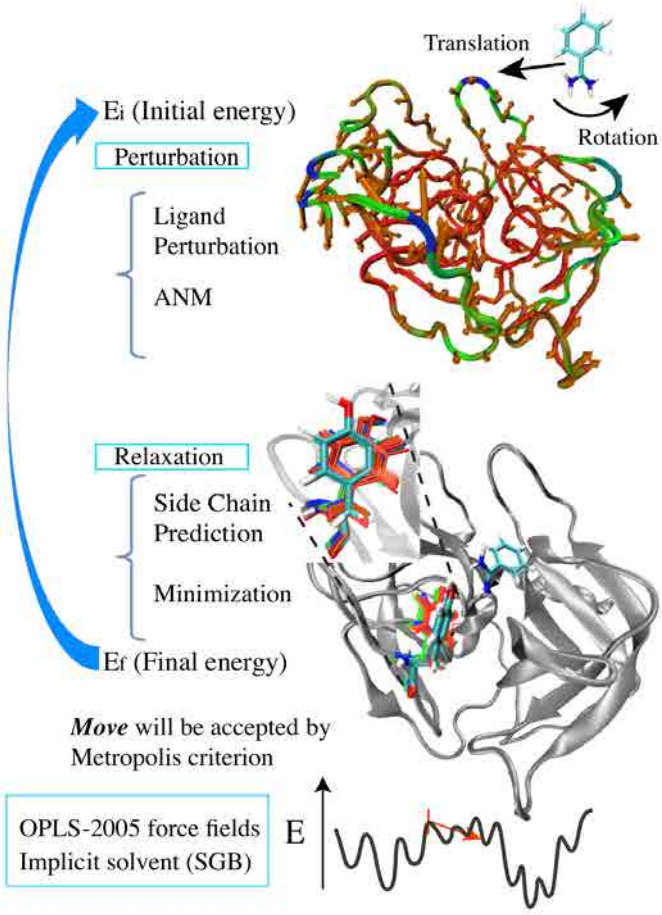


Figure 9. PELE scheme. (https://pele.bsc.es/pele.wt)

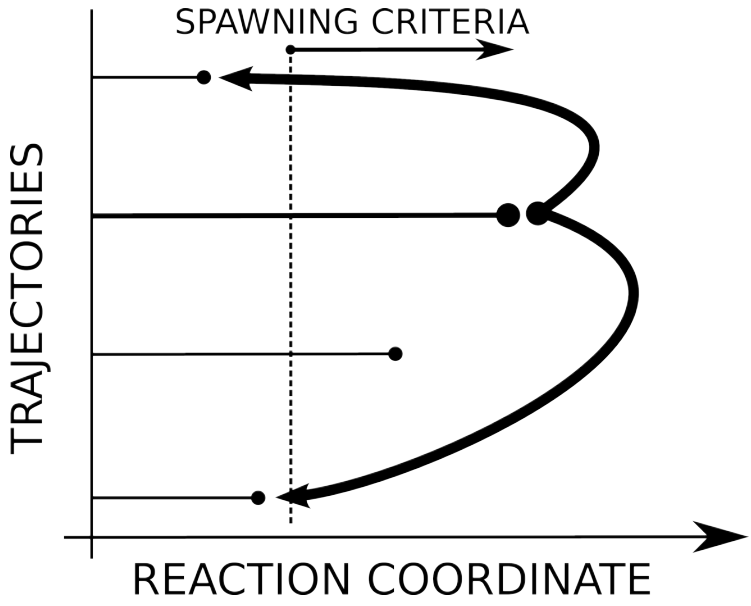


Figure 10. Schematic view of PELE spawning criterion.

1.5.3 PELE applications

PELE algorithm has been applied to study ligand diffusion, induced fit docking, protein local motion and absolute binding free energy estimation. I will proceed to give a detailed explanation of each application.

First PELE application was the study of ligand exit pathways for carbon monoxide in myoglobin, camphor in cytochrome P450cam and palmitic acid in intestinal fatty-acid-binding protein (Borrelli, Vitalis et al. 2005). This study demonstrated the ability to map efficiently microsecond time scale processes consistently with experimental and theoretical data. Also, PELE has been used to study ligand migration of ligand migration in toluene 4-monooxygenase (Hosseini, Brouk et al. 2014).

Induced fit docking between protein and ligands using PELE has demonstrated more accurately induced fit results than commercial softwares. PELE was tested in 88 protein ligand complexes and in 75 % of the cases was able to provide a solution with an RMSD less than 2 A. It shows the accuracy of the sampling procedure and the importance of the all atom physics based potential to identify right conformations when the system is refined (Borrelli, Cossins et al. 2010). Moreover, PELE has been applied to study the substrate binding in enzymology (Hernández-Ortega, Lucas et al. 2011, Hernández-Ortega, Lucas et al. 2012, Hernández‑Ortega, Ferreira et al. 2012).

PELE protein local motion has been tested in two systems: ubiquitin and T4 lysozyme (Cossins, Hosseini et al. 2012). Ubiquitin RMSD and average forces produced by PELE were successfully compared with MD trajectories. Most populated conformations relative to the transition open/close of the T4 lysosome were explored by PELE and validated by MD meta-dynamics.

Absolute binding free energies coming from standard molecular dynamics trajectories can be estimated in combination with Markov State Model (MSM) analysis (Buch, Giorgino et al. 2011). Recently, PELE trajectories have been used in combination with MSM to estimate successful absolute binding free energies from MC trajectories (Takahashi, Gil et al. 2013). Results were in agreement with experimental data and MD simulations indicating PELE capacity for estimation of binding free energies.

Besides these methodological studies, numerous application studies in drug design and enzyme engineering have been published in the recent years.

1.6 Objectives

Over the last few years, computational methods addressing protein/DNA ligand interactions have gained importance in the drug design field due to the increasing computational power, following Moore's Law, and algorithm developments. Scoring methods based on energy functions using force fields and solvent models have been improved leading to a better evaluation of the total energy estimation of the systems per each pose generated. Still, sampling methods are the most expensive part due to the huge number of possible configurations needed for an accurate estimation.

I have focused my thesis on the development and improvement of MC methods to accelerate the exploration of the energy landscape for molecular systems. The three main goals of this thesis are: develop and improve MC algorithms to study: protein/DNA-ligand interactions, protein force response to the stretching and protein-protein docking. In order to do this, the Protein Energy Landscape Exploration (PELE) algorithm was used as the starting point. The following points expand this three main objectives into a more detailed lit

1. We aim to adapt our in-house algorithm PELE to reproduce equivalent DNA conformations than MD simulations for different representative DNA fragments. To this end, we will perform an extensive literature search to improve the force field, the implicit solvent model and the normal mode model for DNA simulations. We will compare PELE conformations with the conformations generated by MD using well-established metrics based on the trajectory analysis.
2. DNA-ligand applications: besides the previous goal, we will apply the new PELE feature to the study of three cisplatin compounds to identify the binding site and the best binder. Moreover, using PELE ligand distribution during the simulation we will estimate the binding free energy of these three compounds and we will compare it with different computational approaches based on MD. We aim to reproduce DNA intercalation process and study the binding energy profile with PELE. To achieve this goal, we will search for an optimum set of parameters in PELE and we will study different algorithm modifications. As a test case, we will study one known intercalator extracted from experimental structures with two different DNA sizes generated canonically with the same DNA intercalation site sequence.

3. Protein-ligand application. We will evaluate the PELE induced fit docking accuracy studying the migration pathway of two protein ligand systems where conformational changes are important for the binding process and we will propose a final bound complex.
4. We aim to measure the system response of a protein to the forced unfolding. To do that, we aim to add an external harmonic force to PELE algorithm to steer or fix selected atoms. We will compare the force-length profiles generated by PELE with steered molecular dynamics simulations.
5. Following the previous goal, we aim to reproduce *in silico* qualitatively the experimental rupture force-length profile for a system proposed by our experimental collaborators generated using the Atomic Force Microscopy technique.
6. Besides these main objectives based on PELE algorithm, we aim to extend a project developed during my master thesis. We aim to develop a novel protocol for a quick generation of protein-protein poses based on a CG model to accelerate the pose generation and discrimination combined with a hydrogen bond network optimization and an energy minimization at all atom level. We aim to validate the protocol with two different protein-protein complexes.

Chapter 2

PELE for DNA-ligands interactions

In the previous chapter, PELE has been introduced, along with the main applications developed in recent years. PELE algorithm has been optimized for protein binding sites where side chains play a significant role in the right binding orientation. DNA structure differs from protein structure due to the chemical difference between residues and nucleobases. DNA double strand is kept due to the hydrogen bonds generated between base pairs. Consequently, it produces lower base mobility in the DNA double helix generating a stable cylindrical shape for small fragments. Thus, side chain prediction algorithm widely used for proteins becomes (to a large degree) neglectable for the ligand-DNA binding process (see Figure 11).

Our first PELE test in DNA focused on fragment simulations to study conformational sampling. For this, we produced a set composed of six representative DNA canonical fragments: BDNA and ADNA with 24, 36 and 48 nucleobases, respectively. We carried out the first test with the small BDNA fragment (24 base pairs), and it showed significant structural artifacts, when compared to the canonical initial strutures, using PELE's protein standard parameters. We found the origin of these artifacts running the different PELE step separately. Even a small initial DNA free minimization was producing a collapse in the double helix reducing DNA volume. This phenomenon was the starting point for an extensive literature search focus on the main parts of the minimization potential energy: force field and solvent model.

First, we searched for an accurate and more recent classical force field since OPLS has not been widely applied to study DNA MD simulations. We found an AMBER force field version called AMBER parmbsc0 (Pérez, Marchán et al. 2007) developed by the Orozco group at IRB. Implementation of this force field in PELE is explained below. Our first PELE minimization with AMBER parmbsc0 in SGBNP (Gallicchio, Zhang et al. 2002) and VDGBNP (Zhu, Shirts et al. 2007) implicit solvents for a 24 bases B-DNA produced a good conformation, similar to the canonical structure. Then, we performed a few PELE steps but hydrogen bond interactions between DNA backbone atoms were overestimated and the minor groove separation collapsed again from 12 Å to 7 Å.

Nevertheless, this could be an expected result because implicit solvent models SGBNP and VDGBNP were developed and tested specifically for proteins. Moreover, comparison of empirical GB models with PB for MD on DNA underlined the importance of the GB model for agreement with PB (Tsui and Case 2000). For this reason, we tested a recent empirical GB solvent called OBC performing an MD simulation with AMBER software (Case, Darden et al. 2012) in a GPU cluster for our canonical B-DNA 24 bases fragment. After 200 ns of simulation, the fragment was stable and with no apparent structural large deviations. We performed the same MD simulation for other five DNA fragments (B-DNA with 36 and 48, A-DNA with 24,36 and 48 bases), and the simulations were also stable for all chains. Due to these evidences, we implemented the OBC implicit solvent in PELE (For details see section 2.2).

Once the force field and solvent were modified, PELE simulations remained stable for long simulations with more than 25000 accepted MC steps in one trajectory (involving ~five days CPU time). Nevertheless, PELE trajectories analysis showed discrepancies between MD and PELE fluctuations due to the ANM default algorithm used. Besides the approximate nature of the ANM modes (driving the backbone motion), this result was also expected because the ANM model parameters were derived again for proteins, where globular shape differs from DNA cylindrical shape. We found a very nice comparison between different ANM models for DNA and RNA (Setny and Zacharias 2013) and we implemented in PELE the most accurate ANM model and parameters (derived from this study). Once the B-DNA 24 bases PELE simulation was satisfactory, we performed additional simulations over the other five DNA fragments of the initial test set.

In the next sections, the new force field, solvent and ANM model implementation will be described in detail. We performed a few test to ensure the accuracy of the implementation comparing our results with small test cases created specifically to this end.

2.1 AMBER parmbsc0 force field

Classical MM force fields introduced in section 1 are widely used for MM simulations such as MD or MC. In particular, AMBER parmbsc0 was specifically developed to carry out long nucleic acids (NA) MD simulations keeping the NA structure parameters stable and reducing the number of artifacts respect to the previous AMBER parm99sb version. Validation test simulations for this force field included two MD DNA simulations of 200 nanoseconds, and 97 individual NA structures studied (Pérez, Marchán et al. 2007). Moreover, it is worthy to

mention that this force field still being improved and a recent version of this force field will be released soon enhancing the parameterization of the NA dihedrals (Ivani 2015).

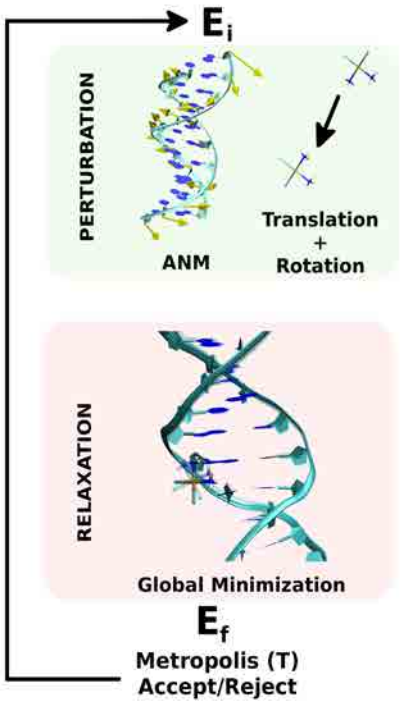


Figure 11. PELE DNA scheme

AMBER Hamiltonian differs from the OPLS one in the 1-4 term scale parameter and the combination rules for the VDW term. OPLS uses a factor 0.5 to scale the electrostatic and VDW interactions for the atom pairs. AMBER uses 0.5 for the VDW term but 0.83 for the electrostatic term. But the rest of the hamiltonian terms remain equal. On the other hand, OPLS uses the geometric average to compute the sigma value of each nonbonding VDW pair term, and AMBER uses the Lorentz-Berthelot rules where radii and epsilon are calculated with the arithmetic and geometric average, respectively. Moreover, it is important to notice that bond, angle and dihedral parameters also differ between force fields in general and between AMBER and OPLS in particular.

PELE uses the impact (Schrödinger) template format where each possible residue/nucleobase has a template. Each template contains the parameters associated with the atoms such as charges, bonds or angles and the parameters related to the residue structure such as zmatrix. A residue/nucleobase located at the beginning or end of the polymer chain has extra atoms making the residue not neutral. Therefore, each residue has three templates per each possible position in the chain: beginning, middle or end.

PELE AMBER parmbsc0 template generation was performed using TINKER (Ponder 2004) molecular package as a reference to extract AMBER parm99sb parameters. It is the closest AMBER force field version to AMBER parmbsc0 available in TINKER. We created a test set with one chain per each residue and nucleobase composed by three repetitions of the residue/nucleobase to generate all the possible chain positions. Generation of these chains was carried out manually with Maestro (2015) software. TINKER uses an intermediate internal format called XYZ where force field atom types are explicitly assigned. Moreover, we used a TINKER option to plot the atom list parameters used for the MM calculation. We used OPLS impact template as the reference and using TINKER output information we replaced OPLS parameters by AMBER parm99sb parameters for each residue/nucleobase created. An automated Python 2.7 script performed everything, and new templates were added to the PELE template folder.

In addition, we generated an automated script to compute and compare energy terms between TINKER and PELE with AMBER parm99sb force field. The script was used to validate the energy of the test set samples for the template generation. We used these small chain samples to include all possible atom pairs inside the cutoff for both codes avoiding problems with differences in nonbonding pair lists. Results showed equal energies for both and verified the templates and hamiltonian modifications implemented in PELE. The successful final test was the energy evaluation and comparison of the small Ubiquitin protein (76 residues) with PDB ID 1UBQ (Vijay-Kumar, Bugg et al. 1987).

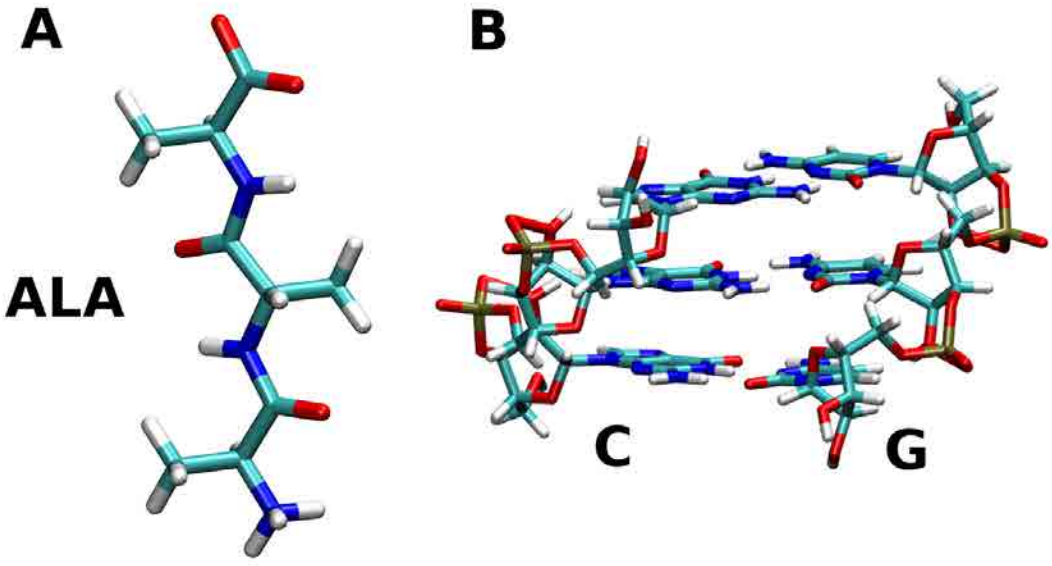
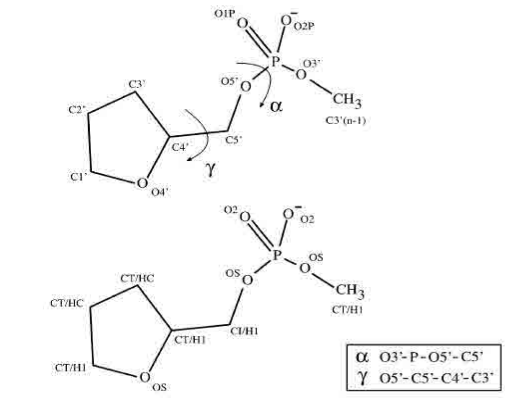


Figure 12. Example of fragments generated to evaluate the PELE energy with AMBER parmbsc0. Panel A, chain composed by three alanine residues. Panel B, three Cytosine-Guanine DNA base pairs.

After that, AMBER parm99sb with bsc0 modifications was implemented modifying manually the new PELE templates extracting the changes directly from Perez et al. 2007 (Pérez, Marchán et al. 2007). In particular, they changed the C5' atom parameterization located in NA backbone to improve the dihedral generated by O3'-P-O5'-C5' and O5'-C5'-C4'-C3' called alpha and gamma, respectively (See Figure 13 A). New atom type CI for C5' added a new parameterization of the torsional terms described in the Figure 13 B.

A



B

Torsion	No. of dihedrals	Vn/2	Phase	Periodicity
X-CI-OS-X	3	1.15	0	3
X-CI-OH-X	3	0.5	0	3
X-CI-CT-X	9	1.4	0	3
CT-OS-CT-CI	1	0.383	0	−3
CT-OS-CT-CI	1	0.1	180	2
H1-CI-CT-OS	1	0.25	0	1
H1-CI-CT-OH	1	0.25	0	1
H1-CT-CI-OS	1	0.25	0	1
H1-CT-CI-OH	1	0.25	0	1
CI-CT-CT-CT	1	0.18	0	−3
CI-CT-CT-CT	1	0.25	180	−2
CI-CT-CT-CT	1	0.2	180	1
OS-P-OS-CI	**1**	**0.185181**	**31.79508**	**−1**
OS-P-OS-CI	**1**	**1.256531**	**351.9596**	**−2**
OS-P-OS-CI	**1**	**0.354858**	**357.24748**	**3**
OH-P-OS-CI	**1**	**0.185181**	**31.79508**	**−1**
OH-P-OS-CI	**1**	**1.256531**	**351.9596**	**−2**
OH-P-OS-CI	**1**	**0.354858**	**357.24748**	**3**
CT-CT-CI-OS	**1**	**1.17804**	**190.97653**	**−1**
CT-CT-CI-OS	**1**	**0.092102**	**295.63279**	**−2**
CT-CT-CI-OS	**1**	**0.96283**	**348.09535**	**3**
CT-CT-CI-OH	**1**	**1.17804**	**190.97653**	**−1**
CT-CT-CI-OH	**1**	**0.092102**	**295.63279**	**−2**
CT-CT-CI-OH	**1**	**0.96283**	**348.09535**	**3**

Figure 13. Panel A, schematic view of the alpha and gamma dihedrals and AMBER atom type definitions. Panel B, table of the new BSC0 torsional parameters for the new atom type CI (C5'). Extracted directly from (Pérez, Marchán et al. 2007).

2.2 OBC implicit solvent

PELE algorithm includes the solvation free energy term using an implicit solvent model where polar and nonpolar terms are evaluated using the surface generalize Born model SGBNP or its variable dielectric version VDGBNP (Zhu, Shirts et al. 2007). This model computes Born radii estimating the molecular surface exposed to the solvent making the calculation computationally expensive. Moreover, DNA molecule is highly negatively charged due to the phosphate groups present in the nucleobase backbone. Our preliminary tests with DNA fragments showed artifacts produced by strong hydrogen bond interactions between DNA backbone atoms leading DNA structures to collapse the minor groove. For these reasons, a new implicit solvent able to solve these two problems was implemented.

The fastest GB methods avoid the integrals over the spherical volumes using mathematically simplified models. In the pairwise version of GB, the integral is approximated as a sum of

contributions from each atom. An analytical expression for the Born radii can be found if we consider the molecule as a set of non-overlapping spheres (Schaefer and Froemmel 1990) where it is important to take into account possible overlap between atom spheres (Schaefer and Froemmel 1990, Hawkins, Cramer et al. 1995). Pairwise overlapping sphere model produces an overcount in the solute region because it is not taking into account the overlap between three or more atoms. Another approach to solve this problem was proposed by Hawking et al. (Hawkins, Cramer et al. 1995, Hawkins, Cramer et al. 1996) evaluating Born radii as a scaling of the neighboring values of R using empirical corrections. Then, Born radii evaluation takes the form

$$R_i^{-1} = a_i^{-1} - \sum_j H(r_{ij}, S_j, a_j) \tag{5}$$

Where H is a sophisticated function with S_j scaling empirical parameters fitted for each atomic element to experiments or numerical Poisson-Boltzmann solutions. This approach has become popular, and many groups have been fitting the S_j parameters using different training sets (Dudek and Ponder 1995, Hawkins, Cramer et al. 1995, Hawkins, Cramer et al. 1996, Srinivasan, Trevathan et al. 1999). The original study (Hawkins, Cramer et al. 1996) was focused on small molecules but later, a better parameterization for macromolecules called OBC was developed (Onufriev, Bashford et al. 2004). OBC model fixed the Born radii overestimation of the buried atoms adding a hyperbolic tangent and three adjustable parameters alpha, beta and gamma (see equation 6).

$$R_i^{-1} = \tilde{\rho}_i^{-1} - \tilde{\rho}_i^{-1} \tanh(\alpha\psi - \beta\psi^2 + \gamma\psi^3) \tag{6}$$

These empirical models are faster than any other GB model and have been implemented in general MD softwares such as GROMACS (Hess, Kutzner et al. 2008), NAMD (Phillips, Braun et al. 2005) or AMBER (Case, Darden et al. 2012). Several MD studies have been carried out to analyze the accuracy of these GB models for proteins, DNA and ligands, where comparison of the energies predicted with GB and PB were in excellent agreement, within 1-2 kcal/mol (Cheatham III, Srinivasan et al. 1998, Srinivasan, Cheatham et al. 1998, Srinivasan, Miller et al. 1998, Srinivasan, Trevathan et al. 1999, Tsui and Case 2000).

Thus, we decided to add the OBC implicit solvent model to PELE, which, besides being the most accurate DNA implicit solvent, provided additional speed up Born radii evaluation and increase the accuracy respect to the PB solution also in proteins. TINKER package was again

used as the reference to check the parameterization of the overlap scale factor, standard GBSA solvent radii per atom and the solvent energy evaluation. Later, we compared SGBNP, VDGBNP and OBC with PB solutions for three different hydrogen bonds to verify the new solvent model accuracy.

We generated OBC templates for all atoms in residues and nucleobases using the same Protein and DNA fragment test set utilized in the AMBER parmbsc0 templates generation. As in AMBER force field parameterization, we checked each atom assignation with a modified TINKER version where solvent parameters were written to an output file. We created an automatic python script with the same algorithm implemented in TINKER for our OBC templates generation. Final OBC solvent templates were stored in the PELE templates data folder.

Hydrogen bonds produced the main DNA instability during PELE simulations. Hence, we compared PELE solvents against PB using the same test set of Mongan et al. (Mongan, Simmerling et al. 2007). We generated, using the Maestro software, three hydrogen bond systems: arginine-aspartate, asparagine-asparagine and aspartate-serine. Arginine-aspartate was the strongest one because was a composition of two hydrogen bonds between NH1-HH12--OD1 and NH2-HH22--OD2 atoms. Asparagine-asparagine corresponds to a hydrogen bond between ND1-HD21--OD1 and aspartate-serine hydrogen bond is between OG-HG--OD2. Also, we tested the interaction between two standard hydrogen atoms with the alanine-alanine system. All chains were capped adding the neutral acetyl group (ACE) and amide group (NMA) (See Figure 14). The reaction coordinate was the distance between heavy atoms of the hydrogen bond except for alanine-alanine system where it was the distance between O and CB. We generated around 120 snapshots modifying the reaction coordinate value from 0 to 7 Å and keeping the planar angle with the hydrogen atom. We used the software APBS (Baker, Sept et al. 2001) to evaluate the solvent polar term with the PB equation in each frame. APBS software takes as input PQR files, which is a format similar to the PDB format but replacing the occupancy and temperature columns in the atom description by the charge and the VDW radii, respectively.

We adopted the same set of PB parameters for the three systems because they had a similar size. Grid points per processor (dime) were set to 161 in the three spatial directions; Coarse-grained (cglen) grid and fine grid (fglen) were set to 80 and 40, respectively. We used the linearized PB equation with the multiple Debye-Huckel boundary condition. Systems were evaluated with 0.15 M of ionic strength and using a constant dielectric 1 for the solute and 80 for the solvent. For further details about the parameter meaning see (Baker, Sept et al. 2001).

PELE solvents OBC, SGBNP and VDGNP were set up with 0.15 M ionic strength without nonbonding cutoff. We modified PELE to extract electrostatic and VDW contributions in different files to compare with PB solutions. VDW and Coulomb binding energies at vacuum for each system were similar for OPLS and AMBER. For this reason, the main difference in the binding energy comes from the solvent contribution.

Figure 14 shows the binding energy for each test case and solvent models (PB, OBC, SGBNP and VDGBNP). Arginine-aspartate system (panel A) shows a local minimum and maximum in the reference model PB. SGBNP and VDGBNP reproduce a maximum a little bit shifted (0.5 Å) to the minima but, more importantly, the energy is considerably higher in 16 kcal/mol and 13 kcal/mol for SGBNP and VDGBNP, respectively. This shift comes from the VDW radii difference between OPLS and AMBER. On the other hand, OBC was able to reproduce exactly the same local minimum energy and position than PB with the same AMBER force field. This model did not generate the local maximum but it is an expected result (Mongan, Simmerling et al. 2007). PB solvent model generated the local energy maximum in Asparagine-Asparagine and the other solvent models overestimated the value of the local minima. Again, OBC was the closest model to reproduce PB interaction energy. The last hydrogen bond test system was aspartate-serine (Panel D). The local maximum was found by SGBNP and VDGBNP but with higher energies up to 10 kcal/mol and the energy of the local minimum were too high. OBC was in close agreement with PB results for the local minimum. Test system alanine-alanine (Panel C) showed a similar behavior for all the solvent models with a small maximum binding energy difference up to 1 kcal/mol between PB and the rest of solvents. According to our own test results (and in agreement with previous literature studies), OBC provided closer solutions to the PB equation than the rest of solvent models for hydrogen bond binding energies.

In addition to the above tests, we computed the binding energy of a PELE trajectory using PB and the other three solvent models. We used beta-trypsin complex (PDB ID 3PTB), which has the benzamidine ligand bound to trypsin. This system is an “easy” test case because the ligand (and partly the protein) is small and with a few degrees of freedom reducing the computational time. It was used, for example, as the test system developed in our laboratory for optimizing the PELE + MSM methodology (Takahashi, Gil et al. 2013). We extracted 150 frames from a real PELE simulation, where the ligand spends the firsts 90 frames in the bulk solvent and the rest exploring the protein surface. Again, our results showed a better agreement between OBC and PB than the other solvent models. SGBNP and VDGBNP overestimated binding energies in all frames.

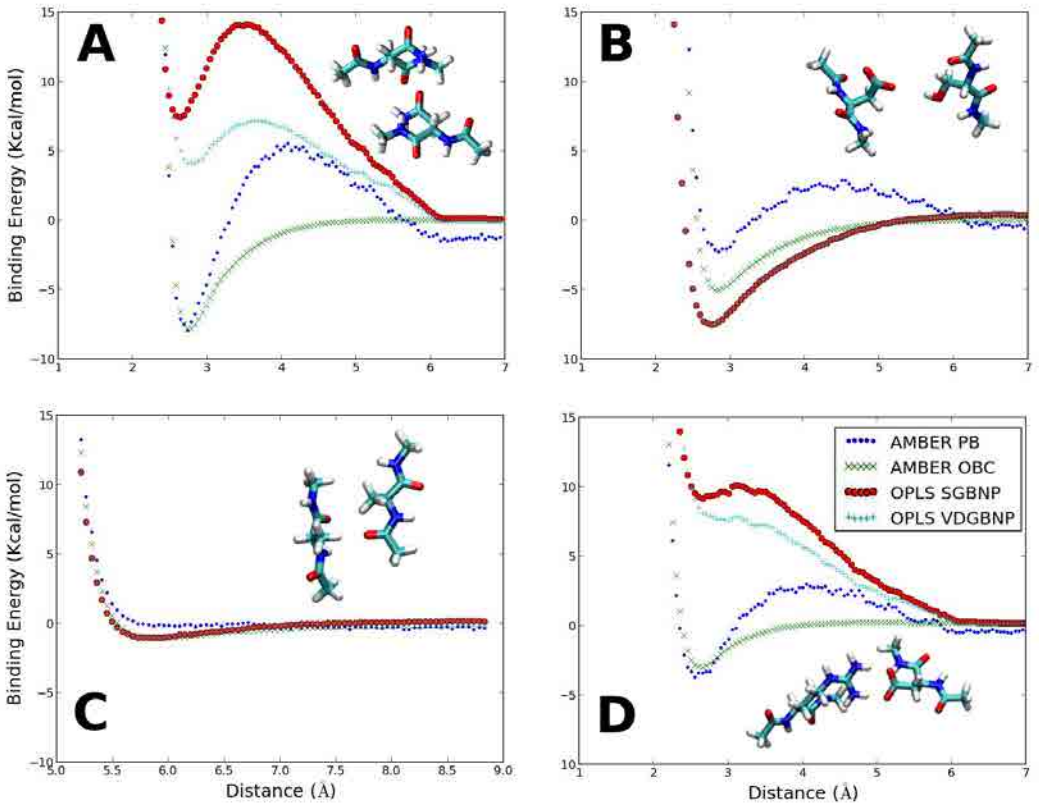


Figure 14. Binding energies along the reaction coordinate for the four systems. Panel A, B, C and D corresponds to the ARG-ASP, ASN-ASN, ALA-ALA and ASP-SER, respectively. Reaction coordinate was chosen as distance between heavy atoms in the hydrogen bonds and distance between CB and O for ALA-ALA system.

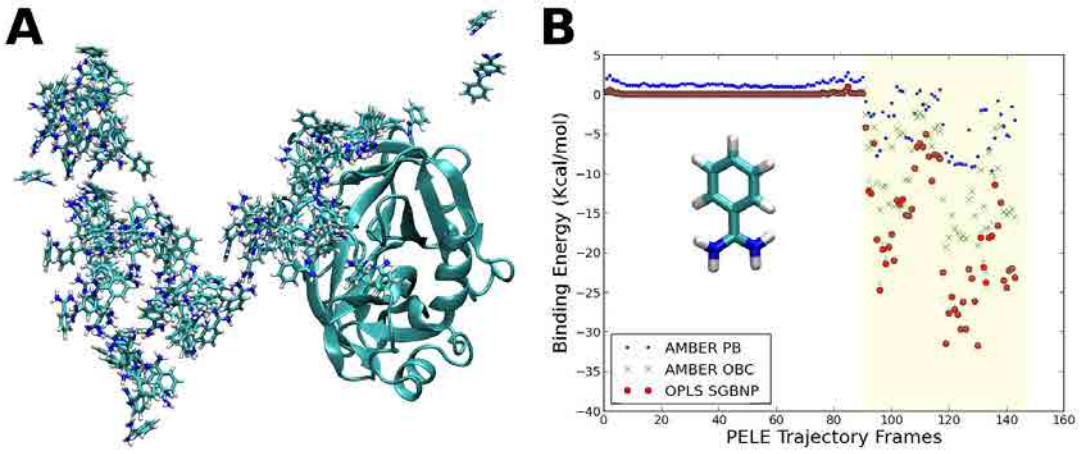


Figure 15. PELE trajectory for 3PTB system (Marquart, Walter et al. 1983). Panel A, visualization of the ligand exploration. Panel B, binding energy plot for each PELE trajectory frame with three solvent approximations. Yellow region shows the frames corresponding to the protein surface ligand exploration. We have removed VDGBNP from the representation because in this case SGBNP produced the same results.

In each PELE iteration, solvent energy must be evaluated and sometimes Born radii must be updated. OBC solvent model uses a faster way to evaluate the Born radii in different situations. PELE simulations comparing OBC with the rest of solvent models produced similar trajectories but saving a 30% of CPU time (data not shown).

2.3 PELE DNA ANM model

PELE ANM model and parameters were originally fitted to reproduce protein fluctuations using alpha carbons as nodes. The DNA double helix structure, on the other site, has a cylindrical shape differing from the standard globular protein shape, making it difficult to translate the ANM parameters. Fortunately, a comparison study was recently published by Setny et al. (Setny and Zacharias 2013) for four elastic network models focus on NA and providing an optimum set of parameters for each one. According to this paper, the best ANM model was the exponential contact (EC) model using the ribose ring center as the node position.

PELE ANM uses atom harmonic constraints in each node to perturb the system. To adapt it to the EC model found, we used as node the C4’ atom. C4' is a DNA backbone atom located in the ribose ring close to its center. EC model computes the force constant needed for the Hessian matrix using equation 7.

$$k_{ij} = k_0 e^{-\left(\frac{r_{ij}}{d}\right)^2} \quad (7)$$

where k_0 and d are fitted parameters with a value of 1.2 kcal/(mol·Å^2) and 5 Å, respectively, and r_{ij} is the node-node distance. We adopted that eigenvectors produced by the 10 smallest modes are used to perturb DNA in order to sample the main global conformations. PELE can update ANM eigenvectors with the new node positions but by default PELE is not updating eigenvectors during a trajectory. The direction of the perturbation is computed as a weighted average over the eigenvectors generated by the ANM model. After each iteration, one random mode is selected and its contribution represents 65% of the final direction. The other 35% contribution comes from the average of the other nine eigenvectors (user adjustable parameters). Once the direction has been estimated, final eigenvectors are scaled by a constant factor of 1.5 and placed in the CG nodes generating the coordinates of a virtual point in the space. A harmonic constraint with zero equilibrium length is generated between each

node atom and the virtual point. Then, a constrained minimization is performed in the DNA molecule to drive the system to a new minimum in this direction.

2.4 PELE DNA conformations test

We propose PELE as a sampling tool independent of the DNA conformation and size. For this reason, A-DNA and B-DNA fragments with 24, 36 and 48 bases were generated in the canonical form as a test set using NAB tool (Case, Cheatham et al. 2005). PDB entry 2K0V corresponding to the (CCTCTGGTCTCC) sequence and its complementary chain was the initial 24 base model. Sequences corresponding to 36 bases and 48 bases were generated replicating the 24 bases sequence motif. Resultant sequences were (CCTCTGGTCTCCCCTCTG) and (CCTCTGGTCTCCCCTCTGGTCTCC) for 36 and 48 bases, respectively. This initial sequence was selected for two main reasons: combines a random distribution of DNA bases pairs and a PDB entry (3LPV (Todd and Lippard 2010)) with the same sequence is available with a cisplatin ligand cross-linked identifying a binding site.

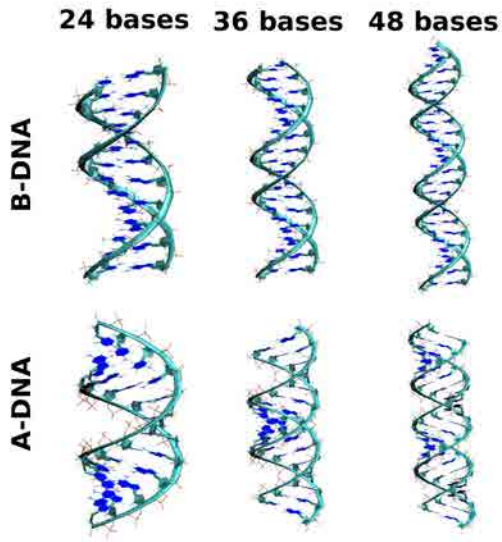


Figure 16. DNA canonical fragments generated with NAB tools for PELE tests.

MD simulations with explicit and implicit solvents of 200 nanoseconds were carried out with each structure to provide a set of DNA conformations. MD simulations were performed using the set up explained in the methods section 2.4.1. Six independent PELE trajectories per DNA fragment were computed to reduce statistical errors associated with the initial conditions. Then, these trajectories were joined in one single trajectory per system removing the first 50 frames of each one, considered part of the equilibration process. All PELE simulations were

carried out using the same set of parameters describes in the below section except the harmonic force constant of the final minimization. Due to limitations in the ANM model (cutoff in connection matrix), there is a linear dependence on the DNA length and this force constant. Long DNA fragments have a larger component of the ANM coefficients at the ends, thus requiring a reduction of the force constant to avoid excessive displacement. Table 1 provides an optimal set of parameters for each DNA conformation and size.

Number of bases	A DNA	B DNA
24	3.0	1.5
36	1.5	0.5
48	1.0	0.0

Table 1. Optimum PELE global minimization force constants (kcal/(mol·Å^2)) for each representative DNA fragment studied.

To quantify the similarities between MD and PELE trajectories we have studied the root mean square fluctuation (RMSF), Principal Components Analysis (PCA) and DNA topological parameters. All these comparison have been performed over the printed conformations generated by both methods along the simulation.

2.4.1 MD protocols

MD simulations have been performed using Amber12 package (Case, Darden et al. 2012). Explicit solvent simulations have been set up using a truncated octahedral water box with TIP3P water molecules (Jorgensen, Chandrasekhar et al. 1983). The distance between the solute unit and the edges of the box was set to 12 Å. The system have been neutralised adding Na+ ions. Force field used to parameterise systems topology was AMBER parmbsc0. The equilibration protocol consists of two minimizations: first just waters followed by the whole system. Then, we performed 200 ps heating up the system to 300 K using a weak-coupling algorithm with constant pressure. The time step used has been 0.5 femtoseconds in the equilibration and production process with the SHAKE (Ryckaert, Ciccotti et al. 1977, Miyamoto and Kollman 1992) algorithm to constrain hydrogen bond lengths. Non-bonding interactions have been evaluated using a cutoff of 9 Å. Particle-Mesh-Ewald (PME) (Darden, York et al. 1993) method has been used to assess long-range electrostatic interactions. Constant pressure and temperature (NPT ensemble) has been applied to the system using a

Berendsen barostat and thermostat (Berendsen, Postma et al. 1984). Checking the convergence of total energy, temperature and pressure, the simulations have been considered equilibrated after one nanosecond. The results shown here have set the time as 0 at the beginning of production process and MD total simulation time have been 200 ns per each DNA fragment studied.

Implicit solvent simulations have been carried out using AMBER parmbsc0 force field and OBC solvent. The equilibration process consists of a global minimization of 500 cycles using implicit solvent and without cutoff. The production process has been performed at 300 K with a time step of 1 fs and without cut off. Total MD simulation length has been 200 ns per each system.

2.4.2 Metrics

DNA conformations produced by PELE and MD along the trajectories have been analysed and compared using the root mean square fluctuation (RMSF), principal component analysis (PCA) and DNA base step parameters. PRODY (Bakan, Meireles et al. 2011) library was used to compute RMSF and PCAs. 3DNA (Lu and Olson 2003) software was used to calculate the DNA topological parameters.

RMSF. Root Mean Square Fluctuation is a measure of the deviation between one atom and a reference position over the conformations sampled during the simulation (See equation 8). We selected the atoms P, C2 and C4' per each DNA base to estimate the RMSF of the DNA fragments, because they are uniformly distributed in similar positions along the DNA bases. RMSF equation is:

$$RMSF = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_0)^2} \quad (8)$$

Where $\vec{r}_i$ are the atomic positions of the atom along the trajectory, $\vec{r}_0$ is the initial position (reference position), and N corresponds to the number of frames.

PCA. Principal Component Analysis (PCA) technique has been used to analyse DNA conformations between PELE and MD. Principal components (PCs) are orthogonal because they are eigenvectors of the variance-covariance matrix. PCs were constructed with an average position of the atoms P, C2 and C4' for each trajectory analysed. To compare trajectories, we have used two metrics based on PCs: PCs inner product matrix and first two

PCs projections per each trajectory frame. Inner product matrix elements are generated using the scalar product between PELE and MD PCs ($I_{ij} = \vec{\mu}_i \cdot \vec{\eta}_j$). They provide a measure of the overlapping between conformations explored by both methods, with a -1 to 1 range because PCs are normalised. As we are interested in the movement direction, matrix components were represented with the absolute value. Projections over PCs were obtained using the scalar product between MD PCs and the displacement vector of each frame respect to the average position of the trajectory (See equation 9. Projection equation is defined as:

$$p_{ij} = \vec{\mu}_i(\vec{x}_j - \langle\vec{x}\rangle) \tag{9}$$

Where i is the eigenvector number, j is the trajectory frame, $\vec{x}$ are the atom coordinates, $\vec{\mu}_i$ is the PC and p_{ij} corresponds to the eigenvector i projection.

Fluctuation DNA analysis. DNA stability study along MD and PELE trajectories was carried out using 3DNA software. We studied all the geometrical DNA parameters provided by 3DNA showing below the base step parameters rise, roll, twist, slide, shift and tilt. We computed the average value along the trajectories but due to the 3DNA limitation managing memory we developed a python script to split the trajectories in multiple small ones and put together the final results for the statistical analysis and graphical representation.

2.4.3 Results

Analyses of the RMSF, PCA and bases topological parameters show that PELE explores an equivalent conformational space as the (200 nanosecond) MD simulations. For the RMSF analysis, we obtained around 5000 frames from each MD and PELE trajectory where the initial structure was chosen as reference for both trajectories. Figure 17 shows the RMSF for representative DNA fragments, where each point indicates the RMSF of three backbone selected atoms P, C2 and C4' and residues are arranged in ascending order from 1 to 24. Initial, medium and final parts of the plot correspond to the 5' and 3'-ends of the double strands. As expected in movements derived from the lowest normal modes, the RMSF plot shows higher fluctuations in the DNA ends for MD than PELE, since these fluctuations correspond to higher frequency modes; all other bases present an excellent agreement.

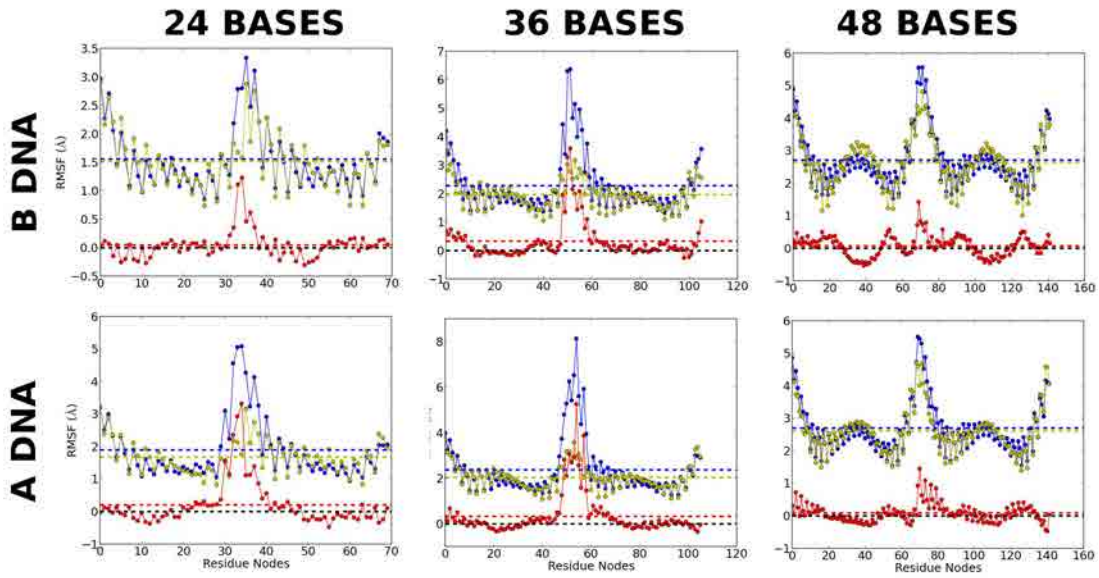


Figure 17. Average RMSF of the six PELE independent simulations. Yellow and blue lines correspond to PELE and MD RMSF, respectively. Colored dashed lines are the mean value per each RMSF plotted. Red line is the RMSF difference between them. Dots represented in each plot are the RMSF value of each P, C4' and C2 atom per each nucleotide base starting from 5' to 3' and 3' to 5'. RMSF scale has been adapted to each plot to show with more detail the differences between PELE and MD.

PCA was used to extract the most important motions from the conformational sampling trajectories. We used the inner product over the first ten principal components (PCs) and projections over the first two PCs to compare both simulation methods. Figure 18 shows the inner product matrix in a colour map with good overlapping between the lowest 4-5 modes, similar to the one we could obtain with two different force fields. As usual, PCs were sorted in decreasing maximum variance order. Thus, the most significant modes are the lowest ones because they represent the highest contribution to the variance of the fluctuations. All six DNA fragments studied showed similar correlations for the inner product matrix diagonal. As expected from applying a simple ANM approximation, in some instances the (variance) ordering from MD and PELE trajectories is shifted.

The first two PCs projections contain the most significant fluctuation information of each trajectory; Figure 19 shows the projections for each fragment simulation for the first two eigenvectors. Clearly, PELE and MD explore the same area showing good agreement between their conformations.

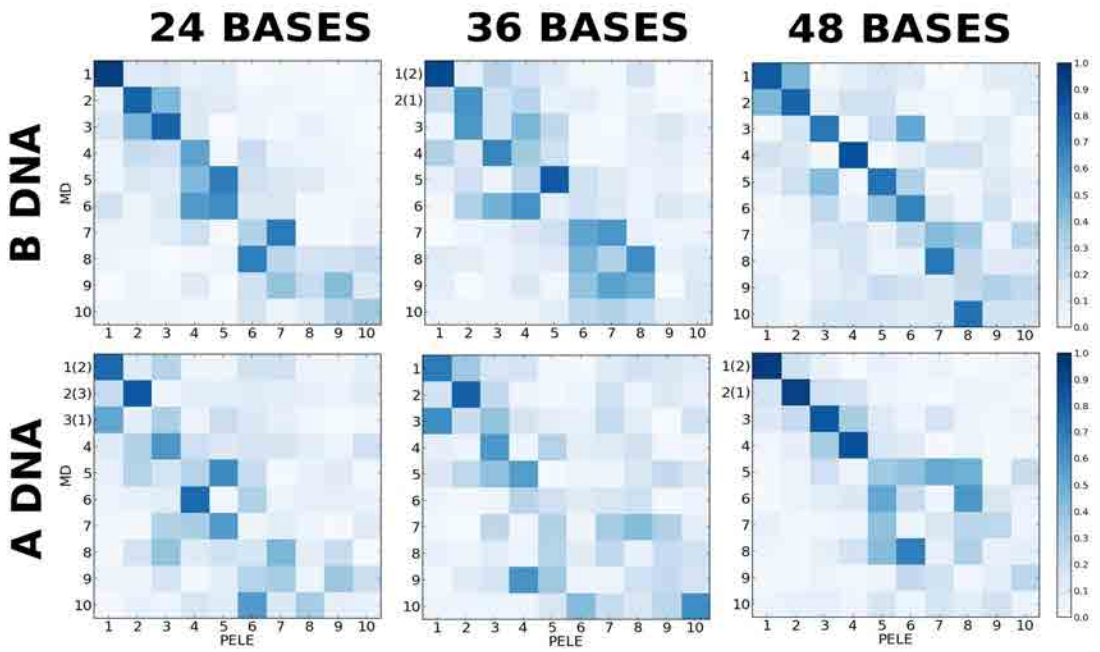


Figure 18. Cross correlation matrix between the lowest ten PCs from the six PELE independent trajectories and MD. All plots have been normalized from -1.0 to 1.0. Correlation or anti-correlation is not important because we are just interested in the module.

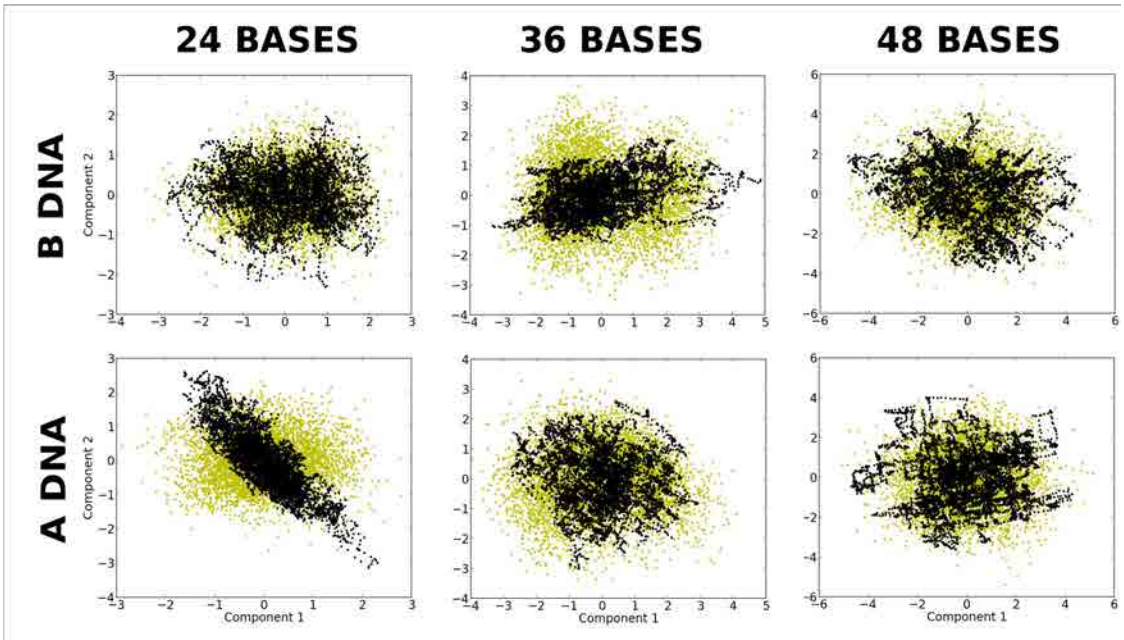


Figure 19. Two dimensional representation of the two lowest projections for each trajectory frame. PELE projections were obtained using the eigenvectors of MD to use the same base in the comparison. Red and blue dots corresponds to MD and PELE, respectively.

Fluctuation analyses of the bases' topological parameters allow us to evaluate the structural integrity along the simulations. We have focused on the parameters: roll, rise, twist, slide, shift and tilt (see 3DNA (Lu and Olson 2003) for a detailed explanation). Figure 20 shows a

comparison between MD trajectories with explicit and implicit solvent and PELE, where the reference value corresponds to the initial structure generated with NAB tools. Overall, the agreement between PELE and MD is excellent. Roll is the only one that showed significant differences between the reference and the simulations, as a result of the strand ends' larger fluctuations. After a few MD and MC steps, DNA's ends were slightly collapsed reducing the chain length and the average roll value. In production runs one might choose to use a weak constraint on the ends if emulating the effects of a larger DNA chain, or to avoid DNA-ligand overestimated interactions (see section 3.2.1.1). In any case, changes in 5´ and 3´ DNA ends do not significantly affect the minor and major groove size and are not important for the binding process.

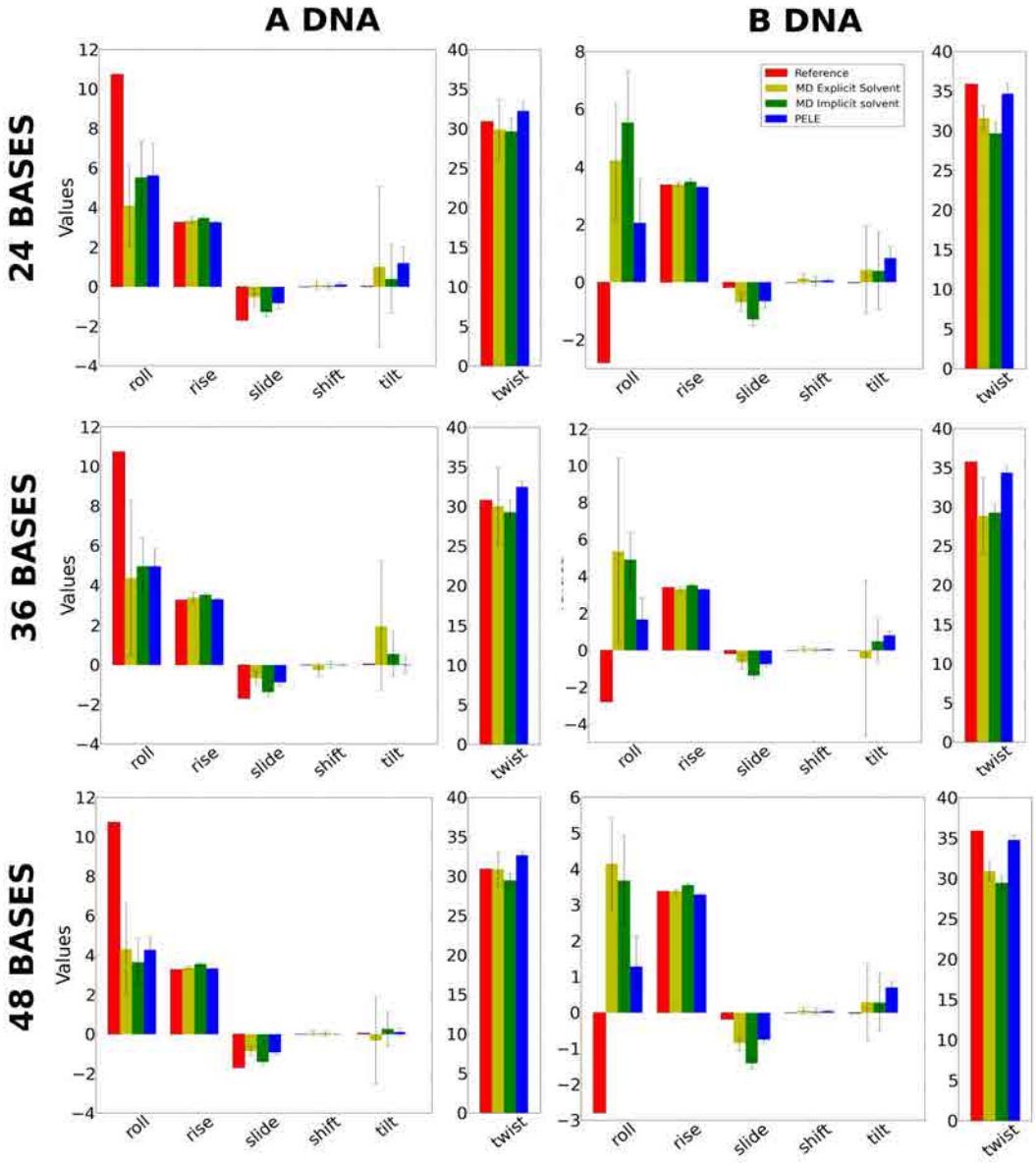


Figure 20. DNA geometric attributes. Roll, rise, twist, slide, shift and tilt base step parameters for the output trajectories generated with PELE, and MD with explicit and implicit solvent. Red bar corresponds to the value of the initial structure used as the reference.

2.4.4 Conclusions

The PELE algorithm is today a well-established Monte Carlo method for studying protein-ligand interactions, with a good compromise between speed and accuracy. Here, we have presented the expansion of the program to allow its usage to study DNA-ligand interactions. To this aim, several modifications including additional implicit solvent, ANM model and a force field have been implemented. All together, with these additions PELE is now able to reproduce conformations obtained at nanosecond scale by MD. In particular, we demonstrated its ability to explore similar DNA conformations obtained with MD for different A-DNA and B-DNA fragments of various sizes. The comparison between DNA structures using RMSF, PCA (inner product and projections) and the base step DNA parameters for both methods confirmed the similarity of the conformational exploration. Certainly, our Monte Carlo based approach has limitations, such as a limited set of normal modes, their approximate nature or the lack of time evolution, that will not make it the best tool for an exhaustive dynamical DNA exploration. Nevertheless, it produces a great quick conformational search to be coupled with ligand dynamics.

Chapter 3

PELE applications

In this chapter, we will show four projects where PELE have been applied to study protein/DNA-ligand interactions. The first project was the study of the porphyrin binding to protein Gun4 combining PELE and MD simulations. The final objective was the study of the ligand-binding mechanism for different protein mutations and two types of porphyrins. Second PELE application consisted of the generation of the first structural model of a bisphosphate substrate bound to human Phosphomannomutase2. We demonstrated that alpha-glucose 1,6-bisphosphate can adopt two low-energy orientations as required for catalysis. For the first time, we applied the OBC implicit solvent to perform a better parameterization of the solvent interaction and speed up each PELE step. The third project was the first PELE DNA application using AMBER parmbsc0 with OBC implicit solvent to study DNA-ligand interactions for three cisplatin drugs. The last project was the study of the binding process of a DNA intercalator using PELE with the spawning criteria.

3.1 Protein-ligand interactions with PELE

3.1.1 Porphyrin binding to Gun4 protein

In oxygenic phototrophs, chlorophylls, hemes and bilins are synthesized by a common branched pathway. Given the phototoxic nature of tetrapyrroles, this pathway must be tightly regulated; an important regulatory role is attributed to ferrochelatase and Mg-chelatase enzymes at the branching between the heme and chlorophyll pathway. Gun4 is a porphyrin-binding protein known to stimulate the Mg-chelatase activity, with obvious potential of regulating the tetrapyrrole pathway, but with no conclusive mechanistic model in the cell. We performed simulations to determine the porphyrin-docking mechanism to Gun4 structures from cyanobacterium Synechocystis 6803. First, we corrected crystallographic loop contacts, which opened the putative binding pocket. Next we determined the binding site for Mg-protoporphyrin IX (MgP) and provided insights on the weaker binding in the W192A mutant as well as the stronger binding in the Gun4-1 variant.

3.1.1.1 Calculations and discussions

Comparison of available crystal structures of Gun4 proteins and the prediction of loops

In order to study the ligand binding mechanism, we first must assess all existing crystallographic structures. Thus, we inspected the three available Gun4 crystal structures deposited in the Protein Data Bank: the Gun4 protein from the cyanobacterium *Synechocystis* PCC 6803 (hereafter *Syn*; (Verdecia, Larkin et al. 2005)), and the Gun4 and so called Gun4-1 mutant (L105F) from the cyanobacterium *Thermosynechococcus elongatus* (hereafter *T.el*; (Davison, Schubert et al. 2005)). As previously reported (Davison, Schubert et al. 2005), *T.el* Gun4 and L105F structures are practically identical except for the L105F change. WT structures, however, show large differences in orientation of α2/α3 and α6/α7 loops (Figure 21), part of the highly-conserved Gun4 core domain (Verdecia, Larkin et al. 2005). Based on an extensive analysis of site-directed *Syn* Gun4 mutant proteins and NMR chemical shift measurements, this core domain was proposed as the porphyrin binding pocket (Verdecia, Larkin et al. 2005) (Figure 21). However, it should be noted that docking of MgP into the binding pocket as proposed by (Verdecia, Larkin et al. 2005) is not possible for *Syn* Gun4. The tight packing derived from the above loop orientations introduces severe steric clashes returning no bound poses from standard docking approaches.

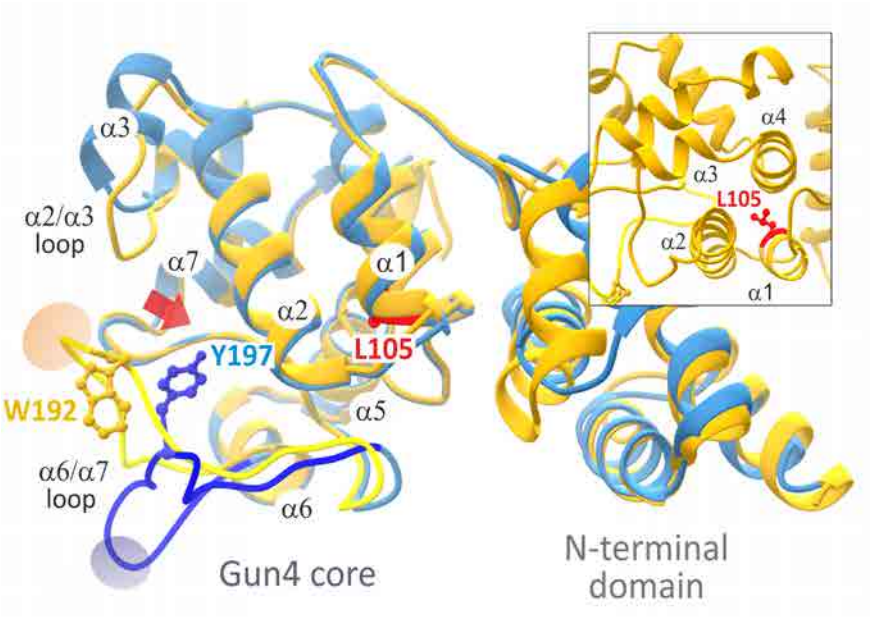


Figure 21. Comparison of two cyanobacterial Gun4 crystal structures. T.el (in blue) and Syn (in gold) Gun4 crystal structures correspond to the PDB codes 1Z3X and 1Y6I, respectively. Loops involved in opening of the binding site are highlighted. Ovals indicate sites of crystal contacts in the symmetric units. The inset shows position of Leu105 (T.el) buried among 1, 2 and 4 helixes.

Interestingly, while the α2/α3 loop has high atomic crystal beta-factors, the α6/α7 one presents quite low numbers, indicating its rigidity (not reasonable in such a large loop) or constrained nature. Inspection of the X-ray symmetric unit indicated critical crystallographic contacts capable of defining (constraining) the α6/α7 loop position. The number of interactions in this region is larger in *Syn* Gun4 and also their nature (salt bridge) is stronger. In particular, in *Syn* we find Trp192-Glu7, Arg191-Glu49 and Thr190-Glu49 contacts, whereas in *T.el* we only observe the Lys192-His69 one. In addition, the difference in these large loop conformations seems to drive a change in the mobile α2/α3 adjacent loop, which further restricts the porphyrin access into the putative binding pocket (Figure 21, Figure 22).

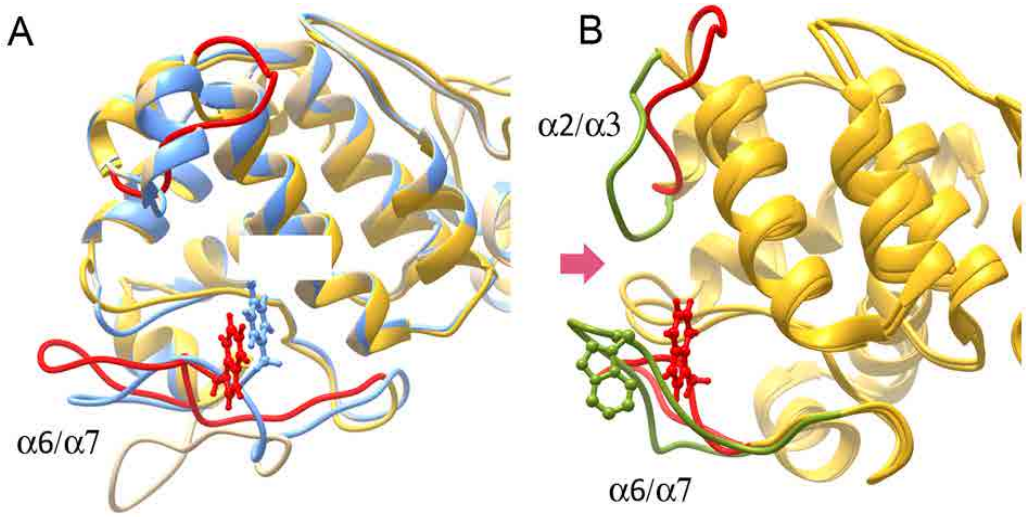


Figure 22. An overview of Gun4 structures with predicted loops. Panel A, alignment of the original T.el structure (in beige) and the Syn (in gold) and T.el (in blue) structures after loop predictions by Prime (Jacobson, Pincus et al. 2004). Note the similar orientation of Trp192 and Tyr197 residues and compare it with Fig. 21. Panel B, The original loop orientation in the Syn Gun4 structure (green color) compared with the lowest energy predictions by Prime without crystal mates (in red). The α2/α3 loop corresponds to Pro122-Phe132 residues and the highlighted part of the α6/α7 loop corresponds to Ser185-Gly195. The Trp192 residue is shown as well as the putative entrance for the porphyrin into the binding pocket (red arrow).

To check if these contacts introduce crystallographic artifacts in the loops conformations, we performed loop prediction in presence and absence of crystal mates. The Prime loop prediction tool (Jacobson, Pincus et al. 2004) was used to model conformations of α2/α3 and α6/α7 in the Gun4 core domain. Due to the loops size, we used the extended loop prediction options. Side chains with less than 7.5 Å of separation were also refined and the 10 lowest energy structures were minimized. Dielectric constant was 1 for internal and 80 for external.

In solution, where no crystal mates are included, *T.el* Gun4 predicted loop is more closed, but still similar to the original structure with an overall open conformation. The *Syn* structure, however, drastically changes from the crystal one (Figure 22A), with an overall alpha carbon

RMSD of 6.1 Å and a similar open structure as the one obtained from *T.el*. Additionally, the α2/α3 loop also significantly moved in *Syn*, further opening the cavity (Figure 22B). We also noted that the conserved aromatic Trp192 residue, located at the α6/α7 loop, markedly turned and adopted a similar position to the homologous Tyr197 residue in *T.el* Gun4. The Trp192/Tyr197 residue has been reported to be important for porphyrin binding (Davison, Schubert et al. 2005) and therefore we selected a Trp192 mutant Gun4 for *in silico*, *in vitro* and *in vivo* analysis (these experimental studies performed by our collaborators at the University of South Bohemia).

The Prime loop prediction software allows running simulations where it takes into account the crystallographic symmetry by placing neighbor chains. In such case, and if neighbor chains help locking loop structures, one would expect to reproduce the crystal structure. This is actually what happens in Gun4, where simulations with crystal mates produce loops with a ~1-2 Å alpha carbon RMSD to the crystal ones (hardly distinguishable from the experimental structures). All together, these results point to a strong bias in the loop conformation by means of crystallographic contacts in both systems. The presence of crystal artifacts, and its study/correction with *in silico* methods, is today a well-established practice (Guallar, Jacobson et al. 2004). More importantly, in *Syn* Gun4 this bias produces a conformation with the loop restricting the access to the expected binding pocket. Thus, we will adopt the corrected semi-open loop conformation, as modeled in solution by Prime, for the next round of simulations.

PELE simulations of porphyrin binding into WT and mutant Syn Gun4 proteins

To map the porphyrin binding mechanism, we performed PELE simulations where the ligand, starting in the bulk solvent, is asked to enter the binding site. Using the spawning algorithms in PELE, the ligand was then asked to enter the active site. The alpha carbon in Phe160 was used as the spawning center (representative of an active site position). This algorithm constrained the ligand random search in a sphere around 18 Å from the spawning center atom. PELE was able to find binding poses in ~24 hours using 48 trajectories (with one trajectory per computing core), using random translations and rotations in the 1-7 Å and 0-90˚ ranges, respectively. Once the ligand reached the active site, local refinement explorations used a smaller spawning sphere, 10 Å, and lower translations of 0.5 Å to better explore the binding energy profile in the bound region. 240 trajectories times 24 hours were used in the refining process for a total of ~50,000 accepted steps.

We focused on the *Syn* Gun4 structure since a detailed analysis of site-directed mutants has been already performed together with NMR measurements (Verdecia, Larkin et al. 2005). In addition, *Syn* Gun4 mutants can be explored both in vitro and in vivo systems. Figure 23 shows four different snapshots underlining the porphyrin docking mechanism into *Syn* Gun4 as observed in PELE's simulation. In panel A we show the initial structure where we placed the porphyrin ligand in the bulk solvent outside Gun4's binding pocket. Panel B shows a recurrent protein surface pre-docking pose observed in both metal and non-metal porphyrins. In this pose, the porphyrin stacks onto Arg113 side chain and forms hydrogen bonds with two glutamines present in the α2/α3 loop. From this pre-docking site, the porphyrin moves into the binding pocket forming an interesting iron axial coordination-like motif with Asn211 and Arg113; we named this site binding pocket-A (Figure 23C and Figure 25A). As shown below, this structure represents a steady bound minimum. From this pose, however, the system can migrate deeper into the binding pocket with less ligand exposure to solvent, and with porphyrin's propionate groups anchored by interactions with Arg214 and Asn211 (Figure 23D and Figure 25B).

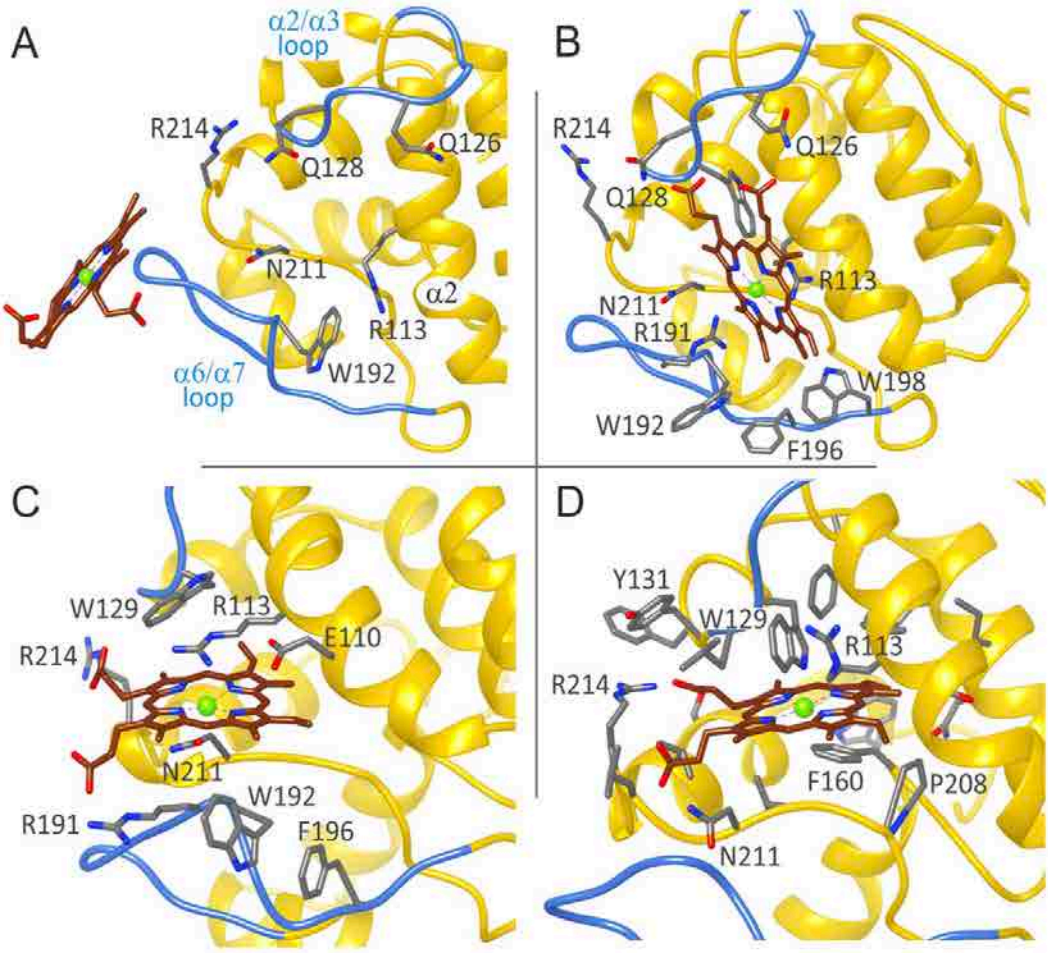


Figure 23. Four snapshots along the porphyrin migration pathway and binding site entrance in Syn Gun4. Panel A, initial structure with a bulk solvent exposed ligand. Panel B, pre-docking pose with stacking onto Arg113 and

hydrogen bonds with Gln126 and Gln128 in the 2/3 loop. Panel C, pocket-A bound structure with an iron axial coordination-like motif with Asn211 and Arg113. Panel D, pocket-B bound deeper structure.

To better characterize the porphyrin binding mode, we performed a PELE refinement search where MgP, starting at pocket-A, is allowed to explore up to 10 Å distance (measured as the ligand center of mass displacement to Asn211 side chain oxygen). Figure 24A shows the protein-MgP interaction energies for the 50000 configurations sampled in this refinement procedure. We clearly observe the existence of two minima, which correspond to binding pockets A and B, with similar (~degenerate) interaction energies. We note that these interaction energies are derived from a classical force field not including any metal coordination component (beyond electrostatic attraction).

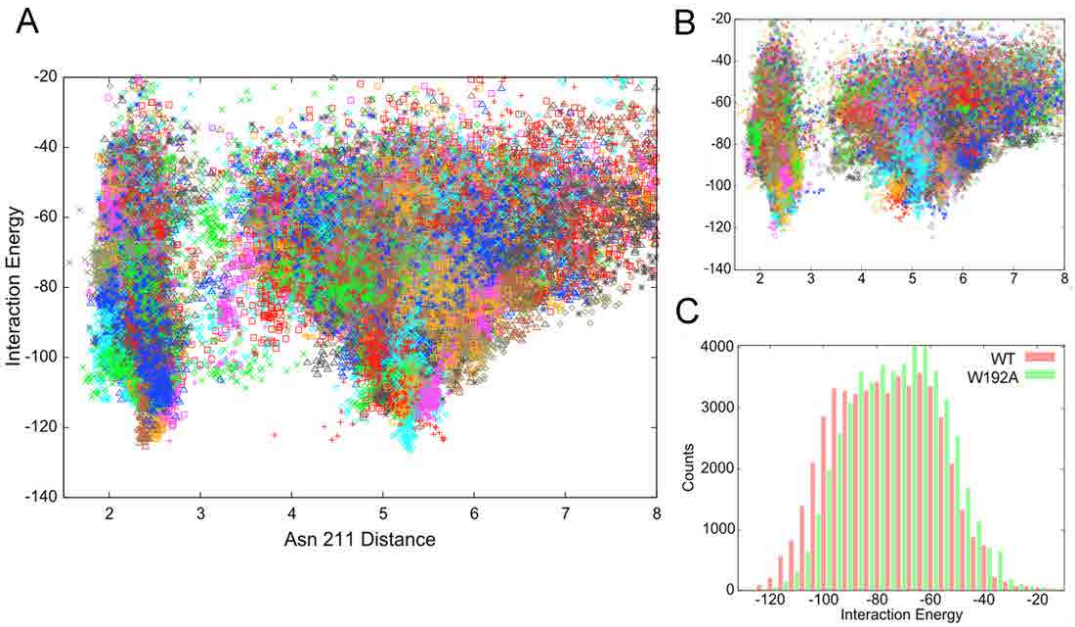


Figure 24. Analyses of the protein-MgP interaction energies by PELE. Panel A, protein-ligand interaction energy against the distance to Asn211 (side chain oxygen atom) along the refinement process for the Syn Gun4. Panel B, the same plot but for the W192A mutant. Panel C, the result of a 4 kcal/mol binning of the interaction energies for the Gun4 (red) and the W192A (green) mutant protein.

To distinguish which of these two minima might better represent the biological system, we compared interacting residues with previously reported studies of side-directed Gun4 mutants (Davison, Schubert et al. 2005, Verdecia, Larkin et al. 2005). Indeed, a number of residues which effect binding are involved in both positions (Figure 25). However, a mutation in Asn211, which is placed in a coordination position with MgP in the less-deeper pocket-A (Figure 25A) has been shown to significantly affect more the affinity to MgP analogue Mg-deuteroporphyrin IX than to deuteroporphyrin IX (D_{IX};(Davison, Schubert et al. 2005). Additionally, the *T.el* homologous residue (Tyr197) to the Trp192, which participates directly in binding only in the pocket-A (Figure 25A), is required for high affinity to porphyrins

(Kopecna, Cabeza de Vaca et al. 2015). On the other hand, replacement of Ser221 has almost zero effect on binding (Verdecia, Larkin et al. 2005), which is not in agreement with the situation in pocket-B, where Ser221 forms a hydrogen bond to MgP (Figure 25B). These are strong indications that the less-deeper docking position-A is closer to the real binding pocket.

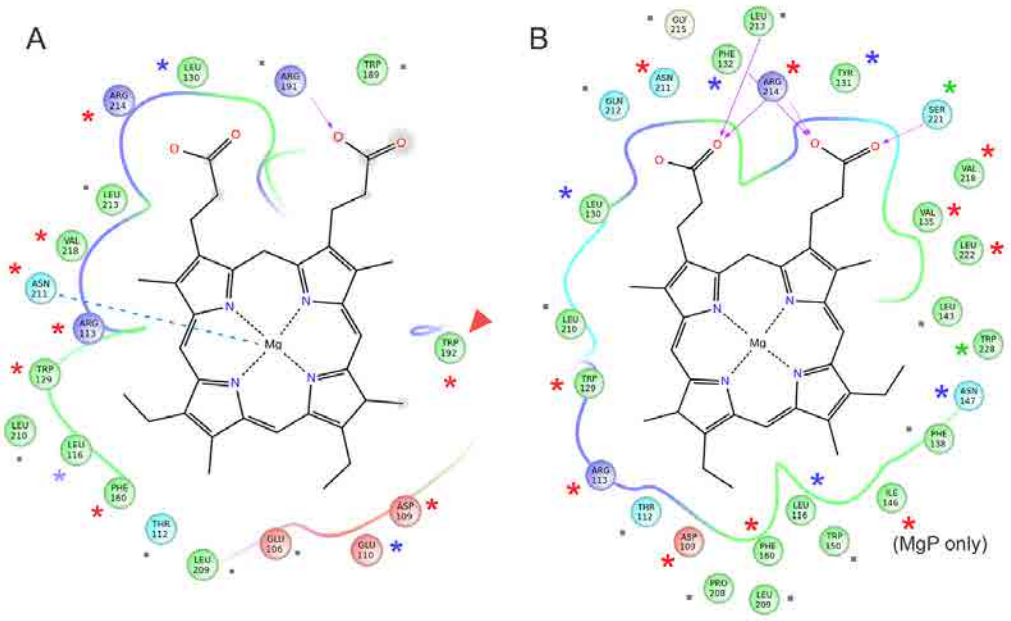


Figure 25. Interaction between MgP and Gun4 residues in two putative binding pockets identified by the PELE simulation. Diagrams panel A and panel B correspond to the position of MgP depicted in Fig. 23C and Fig. 23D, respectively. Line discontinuities in the vicinity of the propionate groups indicate larger exposure to solvent. Arrows represent hydrogen bonds, dashed line in (A) highlights co-ordination of MgP by Asn211 residue and red arrowhead shows Trp192 residue selected for mutagenesis. Residues marked by asterisk were analyzed by site-directed mutagenesis in previous reports (15,16); green asterisks indicate minimal effect of the replacement of given residue on porphyrin binding (an increase of K_d < 2 for MgP), blue means a moderate effect (2-5), red asterisks mark residues, which mutation increased K_d more than five-times. Gray dots mark residues for which no data are available.

To further verify our model we studied the W192A mutant. This residue and Tyr197 are oriented differently in the original *Syn* and *T.el* structures (Figure 21) but, after loop prediction and docking, Trp192 and Tyr197 act similarly. We repeated PELE's local refinement search for W192A (Figure 24B), observing that the mutant has slightly higher interaction energies and less density of points in the bottom of the minima. To better quantify this difference we binned all points in groups of 4 kcal/mol (Figure 24C). By doing so, we can now appreciate more clearly the weaker (shift in) interaction energies for the tryptophan mutant. Such direct role of Trp192 on porphyrin binding is evident from the atomic detailed simulations: inspecting the WT structures along the refinement trajectories, we observe the direct interaction between the tryptophan side chain and the porphyrin group.

Dynamic simulations of T.el WT and L105F Gun4 proteins: cross correlation maps

While the *Syn* Trp192 residue is in direct contact with the bound porphyrin model, Leu105 in *T.el* Gun4 is located far away from the porphyrin-binding site (Figure 21). Moreover, WT and Gun4-1 (L105F) *T.el* crystal structures do not show significant differences (Davison, Schubert et al. 2005). Therefore, the mechanism by which the enigmatic Gun4-1 mutation confers a much tighter porphyrin binding (Davison, Schubert et al. 2005) poses a real challenge. In order to address this issue, and seeking for possible dynamical effects, we turned into MD simulations.

MD simulations were performed with the AMBER11 molecular modelling suit (Case, Darden et al. 2010). All systems were first prepared at pH 7.7 using the Protein Preparation Wizard from Schrodinger (Sastry, Adzhigirey et al. 2013). The parm99 force field was used to define the parameters of the proteins in combination with a truncated octahedron water box containing ~24000 TIP3P water molecules. Na^+ and Cl^- ions were added to neutralize and reach an ionic strength of 0.15 M. After standard equilibration, we performed 200 ns of molecular dynamics at constant pressure and temperature (NPT ensemble) using the Berendsen barostat and thermostat. RMSF analysis was performed using python PRODY library (Bakan, Meireles et al. 2011) and cross-correlations maps were computed using ptraj tool from the Ambertools 12 package (Case and Kollman 2012).

The 200 ns MD simulations of *T.el* Gun4 and Gun4-L105F were analyzed with RMSF (root mean square fluctuation) and cross correlation maps. As seen in Figure 26A, replacement of Leu105 by Phe produces a significant change in the RMSF of several regions up to 1Å. The most important change concerning porphyrin-binding is the clear loss of mobility in the α6/α7 loop (residues 175-215) for the L105F mutant. The (residue movement) difference cross correlation map, shown in Figure 26B, allowed us to establish the mechanism for such a reduction in mobility. If we follow the position around Leu105, we clearly see two main correlated (red) groups (marked in Figure 26B). As expected several residues in contact to Leu105, residues of helix 2 (110-125), are correlated in their motion. More importantly, residues of the α6/α7 loop present marked correlation with Leu105.

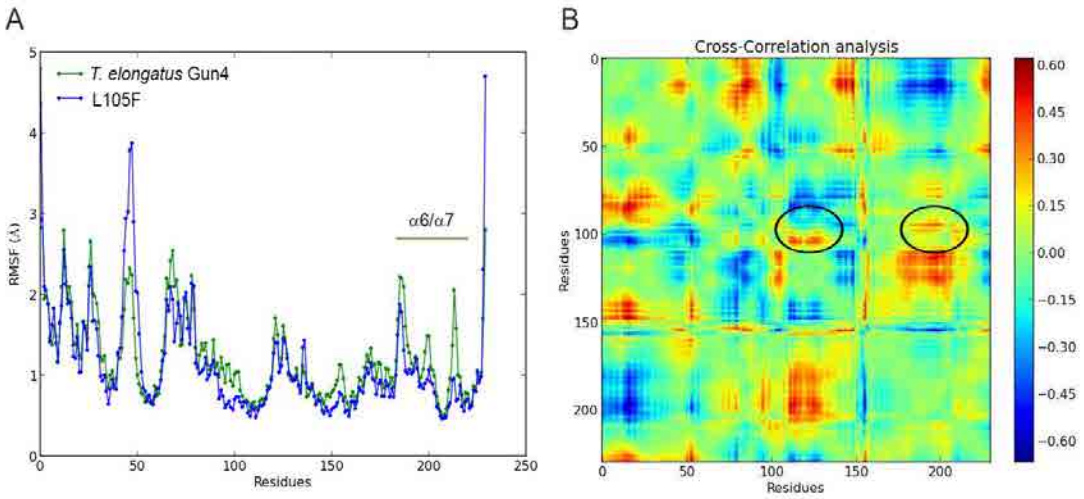


Figure 26. MD analysis of the T.el WT Gun4 and L105F mutant. Panel A, RMSF for both species. Panel B, cross-correlation differences between T.el WT and the L105F mutant. The black circles highlight the region corresponding to helix 2 and the α6/α7 loop.

3.1.1.2 Closure

Our data clearly implies that the orientation of α6/α7 loop is critical for porphyrin binding. After removing the crystallographic artifacts, the α6/α7 loop in both *Syn* and *T.el* Gun4 adopted a similar configuration, indicating that the loop's shape is conserved. Although the original *T.el* structure seems to be in an open conformation (Figure 22), an important residue like Arg196 (equivalent to Arg191 in *Syn* Gun4) is turned away from the cavity, but after loop prediction the Arg191/Arg196 residue as well as the Trp192/Tyr197 residue adopted a similar conformation in both structures. The observed flexibility of the α6/α7 loop and its possible conformation changes after porphyrin binding are consistent with changes in NMR-determined chemical shifts upon addition of D_{IX}. An important finding of our simulation is the ability of Asn211 to co-ordinate MgP, which offers an explanation of the stronger affinity for MgP than protoporphyrin IX (P_{IX}), demonstrated experimentally for both *T.el* and *Syn* proteins (Kopecna, Cabeza de Vaca et al. 2015). As described earlier we expect that the biding pocket-A, which includes co-ordination of MgP by Asn211 (Figure 24A), is closer to the real pocket. Interestingly, the replacement of Trp192 by alanine had a much stronger effect on the binding of MgP than to P_{IX} yielding a Gun4 with almost equal affinity for both these porphyrins (Kopecna, Cabeza de Vaca et al. 2015). It is possible that the Trp192 is critical for the positioning of MgP in the proximity to Asn211. It would be interesting to employ PELE to compare binding energies for MgP and P_{IX} in WT and W192A mutant. However, the lack of accuracy when describing metal interactions by classical force fields does not allow a quantitative comparison between these two porphyrins.

While this project was under review in JBC (accepted in 8th of October for publication), a *Syn* Gun4 structure with bound MgD has been deposited to the Protein Data Bank (Chen, Pu et al. 2015). In this structure the loop orientation is even more opened than predicted by our modelling work and therefore the bound MgD is in a more planar orientation with respect to the α6/α7 loop and it lies closer to Helix 2. However, the main conclusions we derived from the simulations are valid: the α6/α7 loop residues are crucial for the formation of a relatively shallow binding pocket and, importantly, MgP is coordinated by Asn211 and also the Trp192 residue interacts with MgD essentially as we proposed. Asn211 is highly conserved and we speculate that the co-ordination is a universal feature of Gun4 proteins. On the other hand, there might be some variability in orientation of porphyrin propionates groups between cyanobacterial and plant Gun4s. According to (Adhikari, Orler et al. 2009) the homologue Arg214 residue in Arabidopsis Gun4 shows a defect in MgP binding, although it is not essential unlike in the *Syn* counterpart. This result shows the PELE ability to study ligand migration even in complex ligand such as porphyrins.

The level of Gun4-1 protein in *Arabidopsis* is very low (Larkin, Alonso et al. 2003), however this mutation was characterized later using recombinant *T.el* and *Syn* Gun4-1 proteins, showing that both exhibit about ten-times higher affinity for porphyrins than the WT Gun4 (Davison, Schubert et al. 2005). Our results based on MD simulations show a clear steric pathway connecting Leu105 with Helix 2 and the α6/α7 loop (Figure 26), where the loss of mobility is effectively transmitted. Indeed, restricted flexibility of the Helix 2 and the α6/α7 loop can significantly affect the porphyrin binding supporting the critical role of these segments in our docking model (Figure 23).

3.1.2 Conformational Response to Ligand Binding in Phosphomannomutase 2

The most common glycosylation disorder is caused by mutations in the gene encoding phosphomannomutase2 (PMM2), producing a disease still without a cure. PMM2, a homodimer where each chain is composed of two domains, require bis-phosphate sugars, mannose or glucose as activators, opening a possible drug-design path for therapeutic purposes. The crystal structure of human PMM2, however, lacks bound substrate and a key active site loop. In order to speed up drug discovery, we produced the first structural model of a bisphosphate substrate bound to human Phosphomannomutase2. We demonstrated that alpha-glucose 1,6-bisphosphate can adopt two low energy orientations. Upon ligand binding, the two domains come close making the protein more compact, in analogy to the enzyme in the crystals from Leismania Mexicana (PMM_LEIME).

The 2AMY structure deposited in the protein data bank (PDB) represents a good starting point to analyze PMM2 at atomic level. The enzyme is made up by a core (residues 1-81; 189-247) and a cap (residues 86–185) domain, connected by hinge residues. Each domain can be nicely superimposed onto the homologous counterparts seen in PMM1 (PDB: 2FUC, 2FUE) (Silvaggi, Zhang et al. 2006) and in PMM_LEIME, (PDB: 2I54, 2I55) (Kedzierski, Malby et al. 2006). Clearly, in 2AMY the domains adopt an open conformation as expected since no sugar ligands were present during crystallization.

3.1.2.1 Calculations and discussion

The protein data bank 2AMY structure has some problems that need to be fixed before running PELE. The most important aspect is the lack of two Mg2+ ions and the 207-222 loop region. We added the Mg2+ ions and this loop by superposition to the phosphomannomutase1 (PMM1) structure, 2FUC (Silvaggi, Zhang et al. 2006). Importantly, PMM1 has the same loop length and the same initial and final loop residues: PMM1 from Phe215Phe216 to Asp232Phe233 and PMM2 from Phe206Phe207 to Asp223Phe224. Moreover, superposition of the four initial and final alpha carbons, involving residues 206/215, 207/216, 223/232 and 224-233, gives an RMSD of only 0.25Å. Thus, to build our initial model we copied the backbone of the loop in PMM1 into PMM2, and predicted side chains positions with Prime (Jacobson, Pincus et al. 2004). The hydrogen bond network of the initial model was then optimized with the Protein Wizard from Maestro at pH 7 (2015). Five different initial ligand positions were prepared by placing the ligand randomly in the solvent (with a relative solvent accessible area of 1.0) and far away from the active site, with Mg-ligand distances >20 Å.

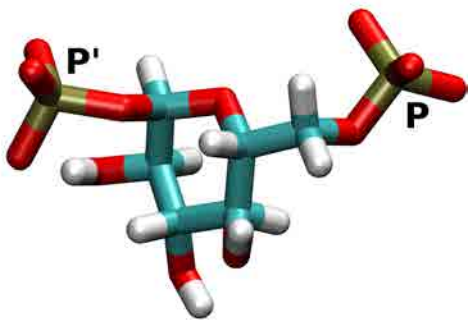


Figure 27. Representation of glucose 1,6-bisphosphate ligand.

Two different PELE exploration runs were performed in this study: a global free search and a local refinement. The global search is performed by combining long (6Å) and short (1.5Å) ligand perturbation steps, with a 75%/25% probability, respectively. Rotations were kept in the [0°:90°] range. Furthermore, a randomly chosen search direction is kept for two MC steps,

allowing a more complete exploration of the entire protein surface. We should emphasize here that no information of the bound structure is used to drive the search. ANM perturbation included the lowest 6 modes, with maximum displacements of the alpha carbon of 1 Å. Within the lowest 6 modes, a randomly chosen mode was kept for 6 steps to facilitate large conformational exploration. The local search used translations of 0.5 Å and rotations in the [0°:180°] range. Furthermore, in order to keep the ligand in the active site, the random search direction was maintained only one iteration.

A reference bound complex was obtained by superposition of our initial model to the holo PMM_LEIME (2I55), and by copying the glucose 1,6-bisphosphate ligand into our initial PMM2 model. Such a reference compound allows us to qualitatively assess the ligand evolution along the free migration performed in PELE. As seen in Figure 27, glucose 1,6-bisphosphate ligand is not completely symmetric and contains two phosphor atoms, identified as P and P', that provide two distinguishable ligand orientations in the active center.

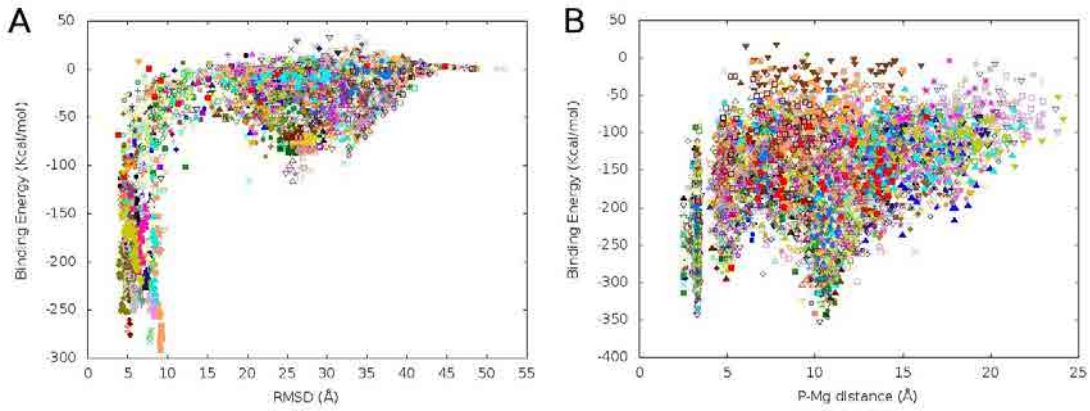


Figure 28. Binding energy profile against the P-Mg distance along the global search (panel A) and the binding site refinement (panel B) processes.

Figure 28A shows the protein-ligand interaction energy against the ligand RMSD, to the bound reference complex, along PELE's 200 independent trajectories global search. As mentioned, the bound reference structure serves only as an indication of the active site area, and not of the exact position. The results clearly indicate that the ligand finds and discriminates the binding site, all other ligand positions on the surface involving significant higher interaction energies (a second surface minimum at ~25-30 Å corresponds to a weak interaction with a second Mg2+ ion present in the 221-226 loop). We should emphasize that the ligand is initially placed far from the active site (~30 Å) and is free to explore, with no bias to the binding site. Interestingly, we find the same number of trajectories where the ligand enters the binding site, interacting with the Mg2+ ion, with either of the two phosphate

atoms, P or P'. Moreover, the energies associated with these two different binding modes are similar.

In order to distinguish the possible preferred binding mode, we proceed by running a ligand refinement search. This refinement step involves small ligand translations and large rotations within the active site, allowing it to reorient but not to move away. Figure 28B shows the protein/ligand interaction energy against the P-Mg distance during the refinement exploration. To further get insight into the preferred binding mode the starting configuration for the 200 refinement trajectories had the same ligand orientation: P' in contact, ~4 Å, with the Mg2+ ion; at this starting conformation the P-Mg distance is ~10 Å. After few Monte Carlo steps, however, the distribution of P-Mg and P'-Mg binding modes reached ~50%. This is seen in Figure 28B where we observe two equally populated energy minima corresponding to both orientations (P-Mg values of ~4 Å and ~10 Å).

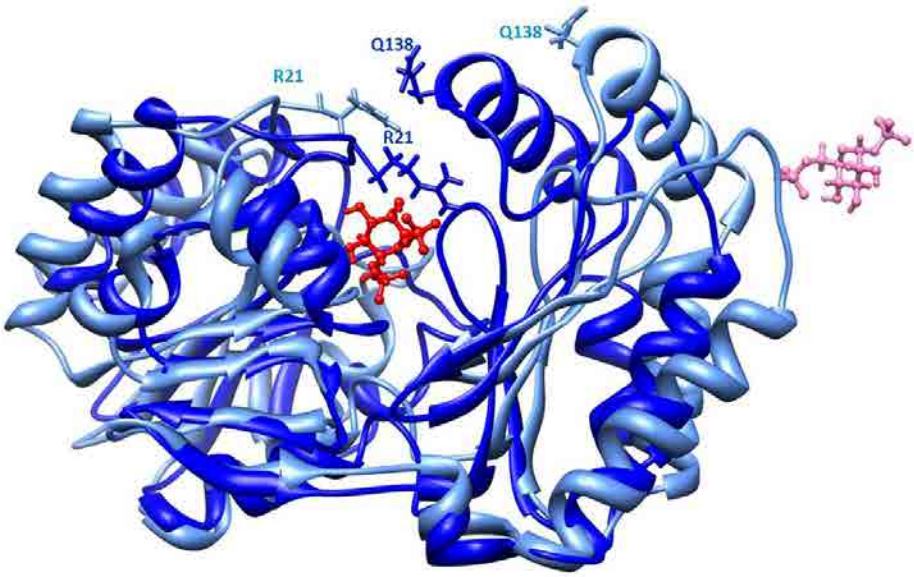


Figure 29. Protein closure along ligand binding. The initial and final positions of the ligand are shown in pink and red, respectively.

PELE's exploration is coupled with large backbone motion, allowing us to monitor closing or opening associated with the ligand dynamics. As mentioned, the initial model obtained from the 2AMY crystal represents an open state. Along with the binding process, however, we clearly observe the closing of both domains around the bound ligand. A static view of domain closure is also shown in Figure 29. The distance between Arg21 and Gln138 alpha carbon experiences the largest change, from 20 to 7.5 Å, moving from open to closed conformations and the exposure of Arg21 to solvent changes dramatically passing from 98% in 2AMY to 15% in the closed model.

Finally, Figure 30 shows the protein-ligand interaction diagrams corresponding to the two binding modes. The orientations of the ligand in the two models are almost symmetrical and can be interconverted by rotation around an axis passing through O5 (i.e. the oxygen in the ring) and C3 (i.e. the carbon opposite to it in the ring). Residues lining the active site pocket can be precisely identified in the closed models whereas they are ill-defined in 2AMY where domains are far apart. While amino acids identity between PMM2 and PMM_LEIME is lower (54.8%) than between the two orthologous human enzymes (63.7%), the active site residues are more conserved with respect PMM_LEIME. Active site residues mostly belong to the core domain (residue 1-81 and 189-247). The two domains come closer upon binding and are bridged by the bis-phosphate sugar: one phosphate bound to Lys189 of the core domain (Lys188 in PMM_LEIME), having the other phosphate bound to Arg134 and Arg141 of the cap domain (Arg133 and Arg140 in PMM_LEIME). A high concentration of positively charged residues is observed in the contact area between domains where Arg21 and Arg141 (Arg19 and Arg140 in PMM_LEIME) should act like clasps as seen in Figure 29.

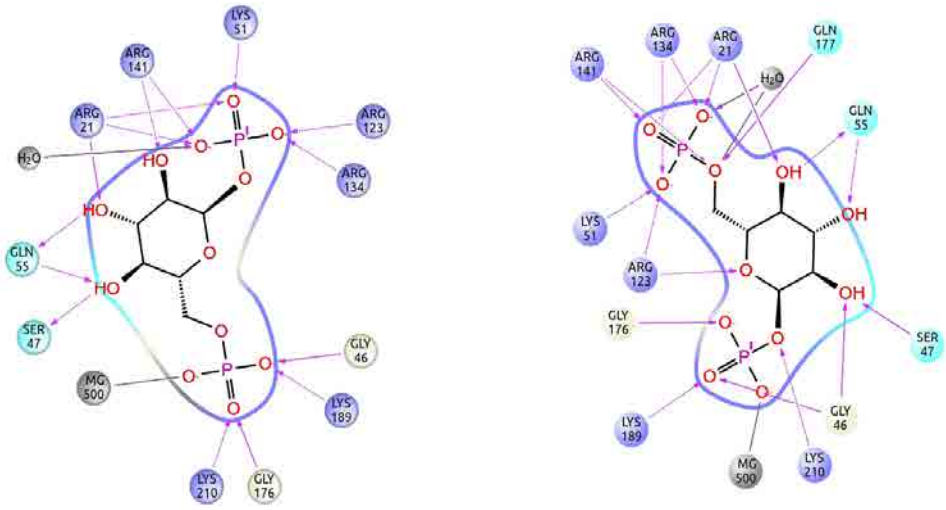


Figure 30. Protein-ligand interaction scheme for the P-Mg and P'-Mg binding modes in PMM2 active site

3.1.2.2 Closure

Our hypothesis states that the current crystallographic structure of PMM2 cannot be used for rational drug discovery because domains are too far apart and Mg2+ ions are missing. Extensive computational and experimental studies (Andreotti, de Vaca et al. 2014) have

confirmed this point. Using our MC protein-ligand sampling techniques, we observed spontaneous binding in the same region as observed in homologous enzymes (PMM_LEIME). The binding process was coupled with significant protein reorganization, adopting a closed structure that required large domain motion.

We have demonstrated that Glc-1,6-P2 can bind in two different modes. Due to the large symmetry of this sugar ligand, we find two equivalent bound structures, where the P or P' phosphates bind to Mg2+. The presence of these two phosphate groups, together with their large negative charge, is crucial for domain closure. Corroborating experimental evidences come from limited proteolysis and thermal shift assay (Andreotti, de Vaca et al. 2014). In fact, the resistance of the protein is minimal with a monophosphate sugar, intermediate with a monophosphate sugar plus vanadate, an inhibitor that mimics phosphate and recreates a complex similar to sugar 1,6-bisphosphate in the active site, and maximal with bis-phosphate sugar. Our models offer insights about the enzyme mechanism and identify active site residues.

3.2 DNA-ligand interactions with PELE

PELE has been widely used for studies of protein-ligand interactions but the methods developed in the chapter two have opened the way to the study of DNA-ligand interactions. In this section, we have investigated the noncovalent binding of cisplatin to DNA analyzing (comparing) microsecond-long unbiased MD and PELE simulations combined with Markov state models (MSMs) to elucidate the thermodynamics of the electrostatic preassociation between cisplatin and DNA. A better understanding of these complex mechanisms, at atomic level, is important to improve platinum-based therapy. The drug, clinically known as cisplatin (Rosenberg, Van Camp et al. 1965, Rosenberg and Vancamp 1969) (cis-diamminedichloro platinum(II)) and its derivates, are among the most widely used antineoplastic agents (Weiss and Christian 1993, Alderden, Hall et al. 2006). In addition to testicular treatment (with more than 95% success rate (Siegel, Naishadham et al. 2012)), these platinum (Pt) compounds have worldwide application in many types of human malignancies (Sherman and Lippard 1987, Boulikas and Vougiouka 2004).

Besides, we report below the first intercalation PELE study of the anticancer drug daunomycin (DIMARCO, Gaetani et al. 1964, Myers and Chabner 1990, Wilhelm, Mukherjee et al. 2012) (also called daunorubicin) with a DNA fragment. We will use two DNA fragments with six and twelve base pairs to test the daunorubicin interaction and how PELE is able to reproduce the opening of the two adjacent base pairs required. As this DNA

conformational change is not reproducible with rigid docking softwares (Gilad and Senderowitz 2013), PELE becomes a valuable tool to study drug design of new intercalators.

3.2.1 Cisplatin drugs

We studied cisplatin association to DNA before covalent binding and how the process diverges from the parent drug, cisplatin, and its hydrolysis products. Figure 31A depicts the studied compounds: CPT for cisplatin, CPT1 for the mono-aquo complex, and CPT2 for the di-aquo complex. Figure 31B shows the crystallographic DNA cisplatin covalently bound structure with the binding site highlighted. We performed extensive MD and PELE simulations for these compounds and the results show unique characteristics for each one. The parent drug exhibits extremely low affinity for DNA thus confirming the unlikelihood of direct binding.

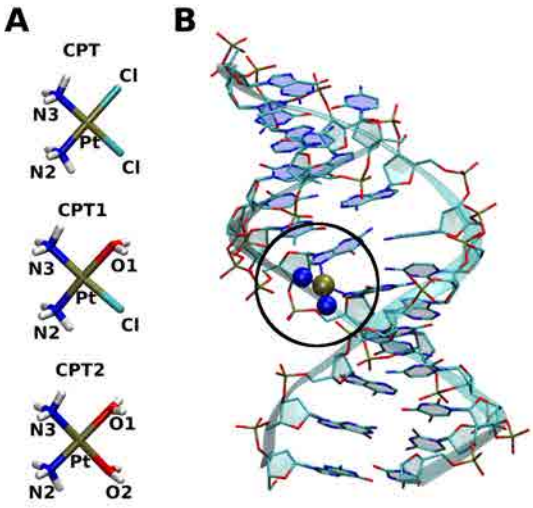


Figure 31. Panel A, view of cisplatin compounds: CPT-parent drug, CPT1-mono-aquo and CPT2-di-aquo. Panel B, representation of PDB ID 3PLV structure showing the cross-linked cisplatin in the binding site (black circle).

3.2.1.1 Calculations and discussion

Molecular dynamics

MD simulation were performed with the PMEMD CUDA module within the AMBER11 molecular modelling suit (Case, Darden et al. 2010). The parmbsc0 force field refinement was used for DNA (Cheatham III, Cieplak et al. 1999, Hornak, Abel et al. 2006, Pérez, Marchán et al. 2007) whereas most parameters for the ligands have been derived through quantum

mechanical calculations. Some parameters were already available either in the literature or in the general AMBER force field (GAFF) (Yao, Plastaras et al. 1994, Wang, Wolf et al. 2004). Waters were incorporated as TIP3P model (Jorgensen, Chandrasekhar et al. 1983). The ligands partial charges, were derived by fitting the electrostatic potential obtained at HF/6-31G(d) level (calculated with Gaussian 03) through the restrained electrostatic potential (RESP) method. The initial DNA structure was taken from the protein database (PDB) entry 2K0V (Bhattacharyya, Ramachandran et al. 2011) corresponding to an undamaged sequence ((CCTCTGGTCTCC)$(GGAGACCAGAGG)). This structure has been chosen because it has an identical sequence to an available crystal structure of a DNA strand containing a cisplatin cross- link (3LPV) (Takahara, Rosenzweig et al. 1995).

We prepared all systems following the same procedure. The DNA + Pt-complexes were neutralized (because we wished to simulate the cell environment with Cl- concentration of ~ 3 mM, no additional salt was added) by addition of the convenient number of Na+ ions and then surrounded by a 15 Å layer of preequilibrated water molecules in a truncated octahedron box containing ~ 45.000 atoms. First, the system was minimized through 10.000 steps: 5000 for ions and water minimization followed by 5000 for the entire system. Then the system's temperature was progressively raised to 300 K using a weak-coupling algorithm during 200 ps of constant pressure dynamics. A time step of 0.5 fs was used throughout the simulations in combination with the SHAKE algorithm to constrain bond lengths involving hydrogen atoms (Ryckaert, Ciccotti et al. 1977, Miyamoto and Kollman 1992). Nonbonded interactions were explicitly evaluated for distances below 9 Å. The particle mesh Ewald method was employed to treat long-range electrostatic interactions (Darden, York et al. 1993). Constant pressure and temperature (NPT ensemble) were maintained by weakly coupling the system to an external bath at 1 bar and 298 K, using the Berendsen barostat and thermostat, respectively (Berendsen, Postma et al. 1984). To improve the extraction of statistical data from the ensemble produced using the Berendsen thermostat we have used a relaxation time of 5 ps (Okumura, Gallicchio et al. 2010). Simulations were considered equilibrated after ~ 1 ns by inspection of convergence of total energy, temperature, and pressure. All computed times presented in this study have as time 0 the beginning of the production process. Total production times are 1200 ns for CPT, 1400 ns for CPT1, and 1100 ns for CPT2, and structures were saved for analysis every 20 ps.

Steered molecular dynamics and binding free energy estimation

SMD performed with the PLUMED (Bonomi, Branduardi et al. 2009) plugin for AMBER molecular dynamics software were used to estimate the binding free energies of the three

compounds under study. The equilibration and production setups for the simulations are identical to the unbiased simulations. The ligand, initially in the active site, is pulled away towards the solvent with an external force applied on the central atom (Pt) of the ligand. The ligand is moved away from the N7 atom of guanine 7 in constant displacements without any particular direction as previous unbiased simulations had shown that the entry/exit routes vary. The spring constant and the velocity were set to 9 kcal/(mol·Å^2) and 0.041 Å/ps, respectively. Binding free energy (ΔF) was estimated from the exponential average work by Jarzynski as $e^{-\beta\Delta F} = \langle e^{-\beta W} \rangle$ where the work (W) for the exponential average corresponds to the work of the pulling external force of each independent trajectory.

Five initial structures for each DNA + ligand system were prepared ensuring that the ligand was in the correct binding position. The systems were then heated and equilibrated in a total of 5 ns. To guarantee that the ligand would not digress from the binding site we have introduced a harmonic constraint of 0.05 kcal/(mol·Å^2) that restrains the ligand to its initial position. This constraint was then removed and the simulations continued for another 1ns where intermediate structures were saved at every 100 ps. At this stage it was necessary to visually inspect the structures to assure that the ligand had not escaped from the binding site. These initial 50 structures were then used for the SMD simulations. The Pt-N7 distance in the initial structures ranges from 3.7 to 4.5 Å. In order to assure that in all simulations the ligand was at the end of the simulations completely in the solution we have set the final position of the ligand to 14 Å and 30 Å. In the case of CPT given the low affinity we find that 14 Å is enough to take the ligand from the active site to the bulk solution while for CPT1 and CPT2 30 Å are required.

Computational details for binding free energy calculations with MMPBSA

The MM-PBSA method was also used to estimate the binding free energies for the three complex systems (Kollman, Massova et al. 2000). The method is applied to the molecular dynamics simulations, where a set of representative structures has been saved. The complete simulation data for each compound were initially aligned to a reference structure of the DNA backbone strand similar to what was done for the MSM studies. The protocol includes a processing of these structures that must initially be stripped of solvent and counterions. The free energy is computed according to the following equation: $\Delta G = \Delta H^{gas} + \Delta G^{solv} - T\Delta S$, where ΔG is the average free energy for the system, and ΔH^{gas} is the average molecular mechanical energy. ΔG^{solv} is the solvation free energy that is obtained by summing the polar (ΔG_{polar}) and nonpolar ($\Delta G_{nonpolar}$) terms. ΔG_{polar} is calculated solving the Poisson-Boltzmann (PB) equation with numerical methods (Sitkoff, Sharp et al. 1994) and $\Delta G_{nonpolar}$

is calculated using the solvent-accessible surface area. The last term, -TΔS, is the solute entropy which has been neglected due to the structural similarity of the studied compounds. This method has been shown to provide a quick and inexpensive manner to estimate binding free energies.

PELE

We started each PELE simulation from six different ligand positions 20 Å far away from the DNA fragment. We used the same B-DNA fragment with 24 bases employed in MD corresponding to the PDB entry 2K0V (Bhattacharyya, Ramachandran et al. 2011). Moreover, this PDB has the same sequence than the undamaged sequence of the cisplatin cross-linked in the G6-G7 base pair 3LPV (Todd and Lippard 2010). This region will be defined as the binding site in our simulation (see Figure 31B).

PELE was set up with an optimal DNA parameters set described in chapter 2. Ligand movement direction was kept during two PELE steps to increase the possibility of the ligand to escape from a local minimum. Translation magnitude was alternated 50 % of the times randomly between 3.0 Å and 1.0 Å and rotation angle was generated with a Gaussian distribution around 72º. This set of parameters was chosen to allow a quick DNA surface exploration with large parameters and a local refinement of the binding regions with small parameters. Two position constraints with 10 kcal/(mol·Å^2) force constant were added into the DNA extreme residues (residue 12 and 24) to avoid artifacts with the ligand interaction due to the finite fragment size. Ligand movement was restricted in a spherical box of 35 Å and the ionic strength of OBC solvent was set to zero as in MD set up. All PELE simulations were carried out with the same input parameters.

Markov state models

Binding free energies were estimated using MSM with the software package EMMA.(Prinz, Wu et al. 2011, Senne, Trendelkamp-Schroer et al. 2012) MSM defines states and uses the transition between them to describe equilibrium properties. PELE and MD simulation frames were aligned to a reference structure using the DNA backbone atoms P, C2 and C4'. MSM was constructed following the next steps (Senne, Trendelkamp-Schroer et al. 2012): 1) extract cartesian coordinates of the central Pt atom of cisplatin molecules; 2) generate 300 clusters using the K-means algorithm; 3) assign each snapshot to a (clustered) microstate using Voronoi discretisation; 4) check connectivity between microstates to determine the largest set of them; 5) assure that the implied timescales become constant after a certain lag time (τ);

stationary distribution of the microstates is computed using $\pi = \pi T_{ij}$ where T_{ij} corresponds to the transition matrix between microstates. After this analysis, the stationary distribution corresponds to the eigenvector with eigenvalue of the transition matrix equal to one. Potential mean force (PMF) profile is then computed using the Boltzmann inversion of the stationary distribution, $G_i = -k_B T \log \pi_i$, and the binding free energy through $\Delta G_0 = -k_B T \log {}^{v_b}/_{v_0} - \Delta w$, where k_B is the Boltzmann constant, T = 300 K, v_0=1661 $Å^3$ (1 M ligand concentration), v_b is the PMF bound volume and Δw corresponds to the difference between the minimum (bound state) and the bulk average (unbound state) values in the PMF profile.

PELE's binding energy (Figure 32**)** with respect to the binding site distance clearly indicates that CPT spends most of the time in the bulk with very few visits to the binding site, where the ligand shows low interaction energies. In fact, other DNA regions (mainly the loose ends) presented better interaction energies than the binding site. Introducing a positive charge into the ligand (by replacing a chlorine substituent by a water molecule) clearly produces a significant shift in the ligand exploration. CPT1 is now able to identify the binding site showing more favourable interaction energies than the neutral variant. Following this trend CPT2 improved the binding site recognition, producing a clear precursor structure for the covalent addition.

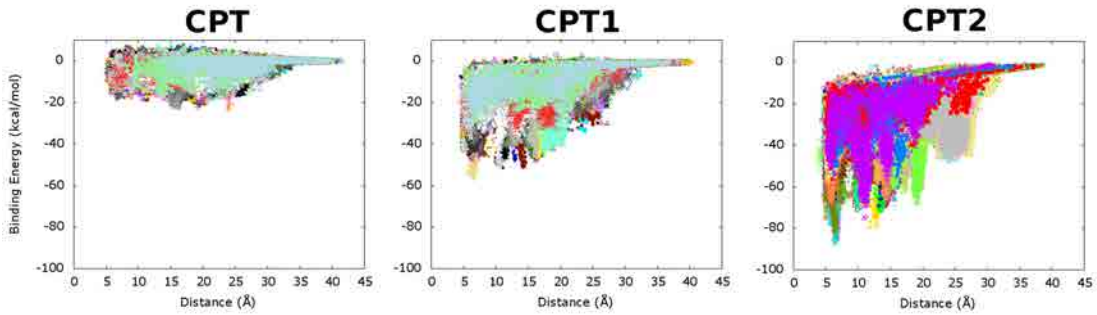


Figure 32. PELE binding energy profiles for CPT, CPT1 and CPT2. Distance is measured from the Pt atom position to the binding site N7 atom from guanine 6.

Figure 33 shows the MD and PELE ligand distribution around the DNA fragment during the simulations for the three ligands, showing excellent agreement between the MC simulation and the microsecond MD. In line with the interaction energy profiles, CPT shows low ligand concentration close to the DNA fragment, CPT1 presents a larger exploration, and CPT2 has the highest cluster concentration around the DNA molecule and especially in the binding region. The ligand structure adopted in both MD and PELE in the binding site is shown in Figure 34, indicating (besides the agreement in both methods) a clear covalent addition

precursor. To further quantify the equivalence of these ligand distributions between PELE and MD, we performed the MSM analyses and computed the absolute binding free energies.

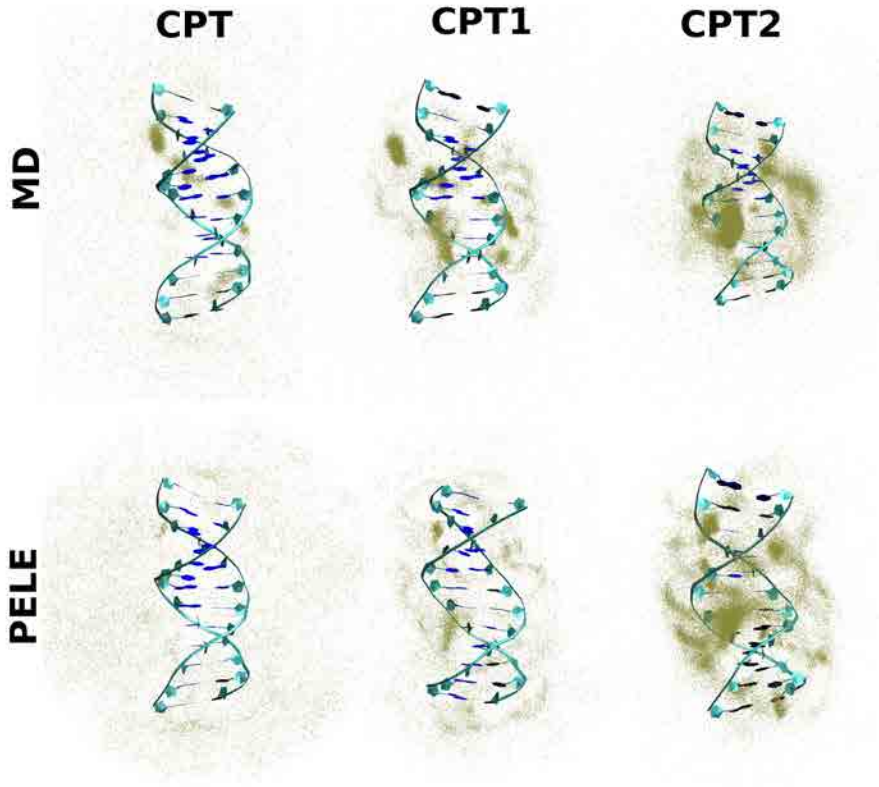


Figure 33. Graphical representation of cisplatin compounds distribution (CPT, CPT1 and CPT2) in MD and PELE trajectories. Yellow dots represent Pt atom position in each trajectory frame, and DNA fragment corresponds to the initial canonical structure.

Obtaining binding free energies is not a trivial task and different approximations such as free energy perturbations (FEP), steered molecular dynamics (SMD), MSM or molecular mechanics Poisson-Boltzmann surface area (MMPBSA) have been implemented. All of them need a massive amount of samples to converge statistically; MD or MC methods were used to generate the system configurations used by these methods. We applied MSM to the long non biased (microsecond) MD trajectories and to PELE's simulations generated in the previous section to further quantify the similarity observed in the ligand distribution. To this end, we have generated the 2D PMF of the MSM and we have estimated binding free energies following the procedure described in the section 3.2.1. Then, we have further compared these binding energies with SMD and MMPBSA to determine the similarity between different methods.

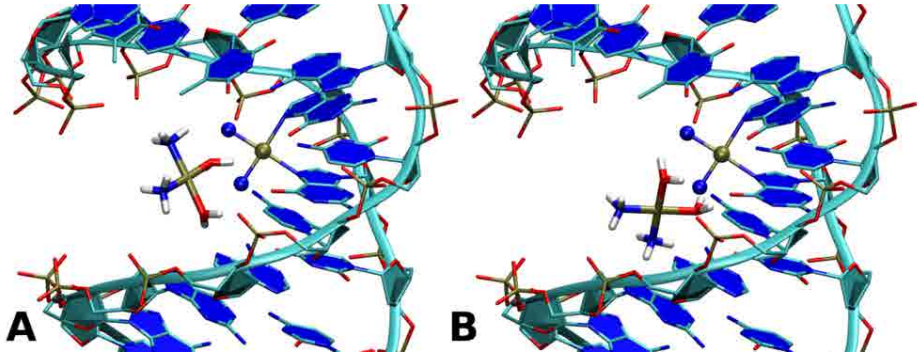


Figure 34. Representative cisplatin di aqua orientations for the binding site cluster. Panel A and B correspond to MD and our MC approach, respectively.

Binding free energies for MSM, SMD and MMPBSA are summarized and compared in Table 2. PELE's results, -0.7±0.2, -2.0±0.5 and -2.8±1.0 kcal/mol for CPT, CPT1 and CPT2, respectively, are in good agreement with those obtained in microsecond MD simulations. The binding free energy for CPT showed the maximum difference (0.7 kcal/mol) between PELE and MD. This difference comes from the ligand conformational sampling: in MD, CPT was able to find more weak local minima around the DNA structure, particularly in the minor groove, increasing the binding free energy. PELE, due to the implicit solvent model and the ligand perturbation, was not able to keep the ligand in these clusters long enough to converge the results. Nevertheless, PELE clearly discriminates ligand affinity and provides quantitative absolute values for ligands with significant binding energy.

		CPT	CPT1	CPT2
	MSM	-1.4	-2.1	-2.8
MD	SMD	-1.6	-2.6	-2.8
	MMPBSA	-2.4	-3.3	-3.8
PELE	MSM	-0.7	-2.0	-2.8

Table 2. Absolute binding free energies (kcal/mol) comparison for CPT, CPT1 and CPT2 drugs.

Finally, we observe very good agreement between MSM and SMD methods. Such an agreement would be expected for our system, a non-buried active site, where pulling the ligand is not coupled to receptor main conformational changes. On the other site, MMPBSA clearly overestimates the absolute binding free energies, a result well documented in other studies, while keeping good relative values.

3.2.1.2 Closure

The potential of PELE in exploring the DNA-ligand conformational space was tested against non-biased microsecond molecular dynamic simulations for cisplatin and two of its aqua derivatives. The well-defined trend (difference) observed for the three ligands (Figure 33), together with the extensive MD simulation data, makes of this system a nice test set. Clearly PELE is capable of reproducing the non-covalent DNA-ligand interactions for the three systems. As expected, differences are only observed for the (very) weak-binding CPT compound, for which the ligand perturbation step introduces too much noise. CPT, however, most likely does not bind to DNA (Lucas, de Vaca et al. 2014); true binders produce quantitative absolute free energies.

Importantly, besides ligand (space) distribution, ranking and absolute free energies, the correct orientation of the pre-covalent bound compound is observed (Figure 34). This is an important feature since obtaining receptor-ligand induced fit orientations is a crucial aspect in drug development projects, from which to design new compounds. Such information is quickly obtained, within 1-2 CPU hours in a commodity cluster (~16 cores), with PELE. Further computation of absolute binding free energies, for instance by using MSM, is not a trivial task, requiring approximately 128 cores for 24 hours. Nevertheless, this still constitutes an improvement over the 1.5 microseconds simulation necessary to reach convergence in MD. Moreover, since each core performs an independent simulation, the method scales linearly with computational resources. This speed up in time and scalability opens the door for *in silico* accurate screening of DNA binders using affordable resources and simulation time.

3.2.2 Intercalators

Intercalation is one of the most frequent DNA binding modes for small aromatic molecules. It consists of the insertion of a small molecule or fragment between two adjacent DNA base pairs with or without other interaction with the grooves. For this insertion, a hydrophobic pocket must be generated with a gap opening between the stacked base pair. In general, intercalation occurs into CG intercalation site (Boer, Canals et al. 2009).

We have used PELE to reproduce the binding process of an intercalator with two DNA different fragments. For this, we used the DNA sequence from the bound complex DNA-daunomycin with PDB ID 1DA9 (Leonard, Hambley et al. 1993) for the first test and a dodecamer generated artificially in the canonical B-DNA conformation for the second test. In

this second fragment, we modified it to include in the center the CpG step (CG base pair followed by a GC base pair in 5'-3' sense), that has shown the strongest binding affinity for daunomycin (Chen, Gresh et al. 1985, Chaires, Herrera et al. 1990, Roche, Thomson et al. 1994).

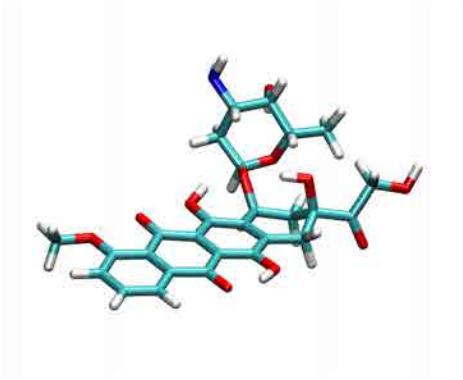


Figure 35. Daunomycin DNA intercalator molecule.

3.2.2.1 Calculations and discussions

Daunomycin molecule (see Figure 35) was extracted from the bound complex PDB ID 1DA9 with Maestro (2015) and prepared (protonated at pH 7.0 and relaxed) using Protein Preparation Wizard (Sastry, Adzhigirey et al. 2013). Daunomycin parameterization was performed using OPLS-AA force field (Jorgensen, Maxwell et al. 1996) and we generated the two B-DNA canonical structures with NAB tool (Case and Kollman 2012). First structure sequence was ACCGGT corresponding to the 1DA9 sequence and second structure was the dodecamer GCGCACGTGCGC.

Simulations were performed with AMBER parmBSC0 and OBC solvent with zero ionic strength where PELE's DNA parameters were the optimal ones described in chapter 2. Ligand movement was restricted to a prismatic box centered in the DNA geometric center with dimensions 45, 45, 31 for X, Y and Z, respectively. The ligand translation was alternated between 1 Å and 3 Å with 40 % of probability and 4 Å with 20 % of probability. The rotation was exchanged between 90 and 18 degrees with 50 % of probability. To enhance the intercalation process, the perturbation random direction was kept for three steps. Furthermore, to increase the search we use a 10 Å spawning criteria based on the maximum distance between N1 atom of the guanine of the binding site and C3 from daunomycin. Maximum overlap factor was set to 60 % to allow intercalated positions of the ligand during the ligand perturbation (which are relaxed in the minimization step). Besides, both DNA systems used the same parameters set.

PELE was able to simulate the intercalation binding process for daunomycin in these two DNA systems. We used a small DNA fragment as a toy system to find the best PELE parameters to optimize the intercalation. For the small DNA fragment, PELE only needed 12 cores and less than an hour to intercalate daunomycin. Figure 36A shows three frames corresponding to the initial, middle and final part of the simulation. The spawning criteria (less than 10 Å) reduced the conformational search space of the ligand trying different orientations close to the binding site. In the small DNA fragment we started with the ligand close to the binding site (with a good orientation) but in the middle step we can see how the ligand is exploring other worst conformations until it finds an intercalation pose. For the large system, we put the ligand far away of the DNA fragment (around 30 Å). As the small system presented the intercalation binding site in an end nucleobases, which is not realistic, we used the large DNA fragment to validate PELE intercalation results in a DNA molecule where intercalation takes place in the middle of a DNA molecule. Figure 36B shows four frames corresponding to the PELE trajectory of intercalation for the large DNA fragment. Due to the DNA fragment size and the initial ligand distance we used 48 cores to speed up the intercalation process. Ligand approximation to the binding site, as a result of electrostatic attraction, induces CpG opening due to VDW interactions allowing ligand entrance into the binding site (See Figure 36B). Obviously this depends on the DNA bases fluctuations and the overall DNA conformation (the combination of both occurrences turning entrance into a rare event); moreover, intercalator proximity reduces DNA fluctuations . Once the ligand enters into the binding site the dispersion forces produced between the aromatic rings of the ligand and the DNA bases stabilize the fluctuations between them allowing just small movements keeping the relative planar orientation.

When ligands are intercalated, binding energy profiles for both systems show a clear energy minimum. Figure 37 displays the binding energy profile for the small and large DNA systems respect to the distance between N1 atom of the guanine of the binding site and C3 from daunomycin; both systems present an energy minimum corresponding to the intercalated structures. The minimum for the large DNA was of ~-50 kcal/mol at 5 Å but in the small DNA fragment case (panel A), however, the minimum is slightly shifted (to 6.5 Å) as a result of the high base pairs flexibility in the chain ends, allowing a small opening of the base pair. In the large system, the rest of DNA bases avoided this opening, keeping the right binding orientation during the PELE simulation.

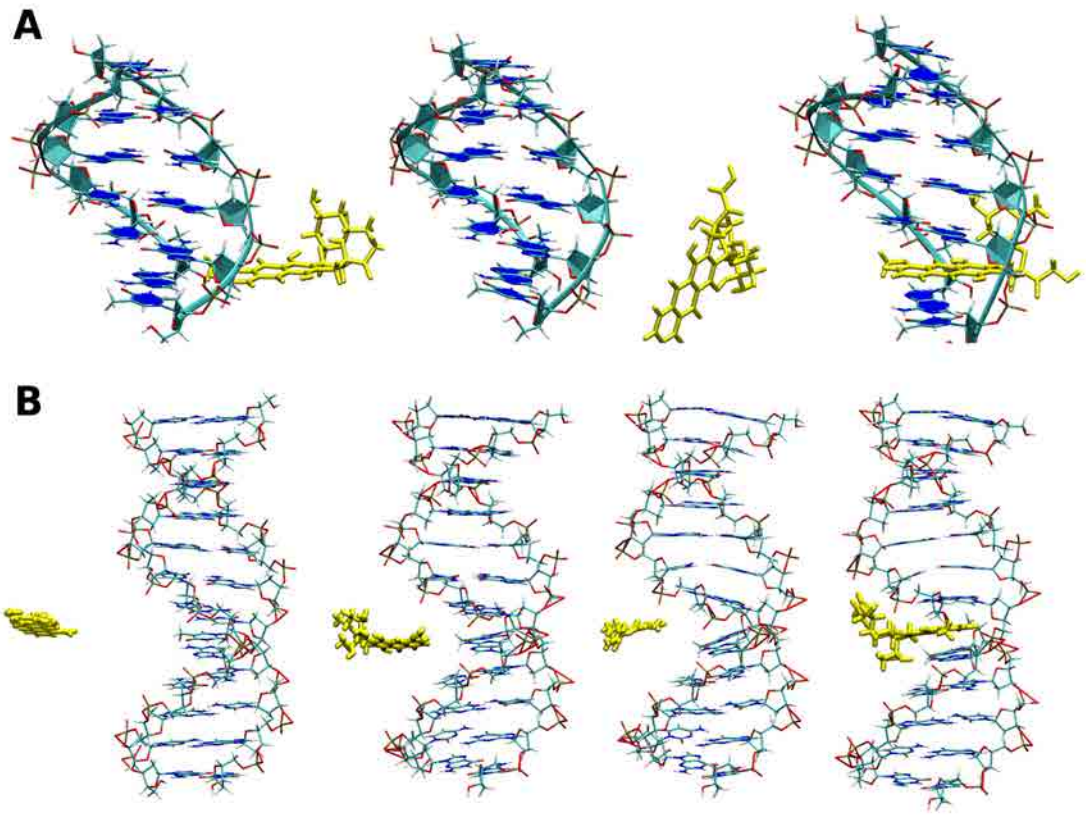


Figure 36. Panel A, PELE trajectory frames corresponding to the accepted steps 0, 191 and 301 (from left to right) for the daunomycin intercalation with the small DNA fragment. Panel B, daunomycin intercalation PELE trajectory frames corresponding to the accepted steps 0, 118, 235 and 354 (from left to right) for the large DNA fragment.

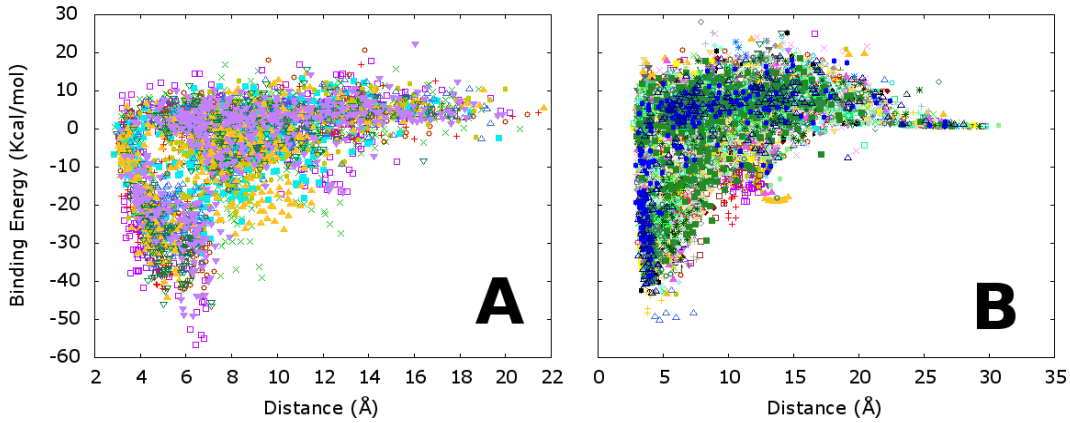


Figure 37. Panel A, binding energy profile corresponding to 1DA9 DNA fragment (ACCGGT). Panel B, binding energy profile for the large DNA sequence (GCGCACGTGCGC). Distance is measured between the N1 atom of the binding site guanine and the C3 from daunomycin.

3.2.2.2 Closure

Intercalated poses cannot be generated with rigid body docking techniques due to the large conformational change needed in the adjacent base pairs. In chapter 2, PELE demonstrated its ability to reproduce DNA conformations for different representative DNA fragments. Now, PELE has shown the ability to simulate the DNA intercalation process for daunomycin ligand with two DNA fragments opening the door for future studies based on intercalators. We have reproduced, with PELE, in few hours results that need large computational resources in MD making PELE a promising tool for these studies.

Chapter 4

Steering Proteins with MC

SMD has been widely used to simulate AFM experiments in order to obtain an atomic description of unfolding force-extension profiles in proteins (Krammer, Lu et al. 1999, Lu and Schulten 1999, Lu and Olson 2003). Due to computer limitations, these simulations stretch the molecule at a speed significantly faster than experiments, resulting in force overestimation (Lu and Schulten 1999, Rico, Gonzalez et al. 2013). Thus, current AFM modelling is based mostly on a qualitative exploration (and agreement with AFM observations (Lu and Schulten 1999)).

MC methods are traditionally seen as an alternative to MD techniques. As described above, our MC approach combined with protein structure prediction tools has shown the capabilities of reproducing protein dynamics at 1-2 orders of magnitude faster rate than MD (Cossins, Hosseini et al. 2012). These technological advances open the possibilities of modelling multiple experiments in a timely manner involving, for example, different initial conditions or pulling residues. For this reason, we have added in PELE a steered particle protocol capable of reproducing AFM experiments. We applied the modified PELE algorithm to study two systems: ubiquitin (Vijay-Kumar, Bugg et al. 1987) and azurin (Nar, Messerschmidt et al. 1992) where we performed a comparison with SMD trajectories to validate and test the methods. Once the methods were validated, we developed a protocol to simulate AFM experiments on apo and holo azurin (with and without the coordinated cupper metal ion) developed by Pau Gorostiza's group at IBEC (Giannotti, Cabeza de Vaca et al. 2015). Moreover, we implemented the steering procedure in MCPRO algorithm (Jorgensen and Tirado–Rives 2005) to estimate the unfolding free energy of deca-alanine molecule.

4.1 MC scheme to stretch molecules

We have added the possibilities of including atom harmonic constraints to a moving virtual bead (VB) in a similar fashion to SMD. Thus, at each MC step the VB is displaced by a fixed amount in the desired direction. The VB starting position is the same as the pulled atom, giving an initial force of zero. Moreover, a fixed (strong) harmonic constraint is added to another atom called fixed atom emulating surface fixation of the substrate in AFM

experiments. Then, at each MC iteration, PELE computes the harmonic force induced by the VB motion, modelling the corresponding force measured by the cantilever (see Figure 38). Due to the third Newton law, force response of the system is equal to the force applied by the harmonic constraint in the pulled atom. Thus, this force can be computed directly using the first derivation of the harmonic potential.

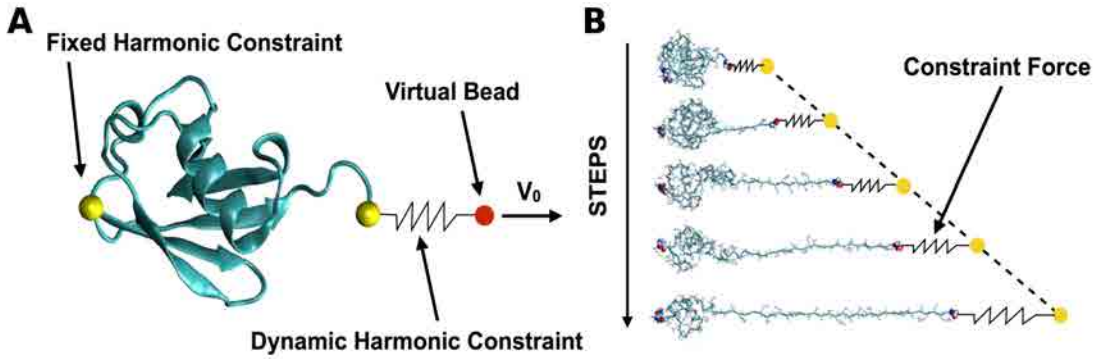


Figure 38. Panel A, MC steering scheme. Panel B, ubiquitin protein unfolding simulated with PELE.

4.1.1 PELE test case I: ubiquitin

Ubiquitin (UBQ) is a small regulatory protein with 76 residues and a molecular mass around 8.5 kDa found in all eukaryotic cells. It performs its functions through conjugation with many target proteins and builds chains due to the linkage possibility between one of the seven lysines and the C-terminal. UBQ was identified in 1975 (Goldstein, Scheid et al. 1975) but the basic functions were elucidated in the early 1980s (Hershko, Eytan et al. 1982). Ubiquitin has been selected as the first test set because it is a well-known small protein where previous studies have compared SMD with AFM experiments (Carrion-Vazquez, Li et al. 2003).

4.1.1.1 Calculations and discussions

PELE simulations have been carried out using the PDB entry 1UBQ (Vijay-Kumar, Bugg et al. 1987) where the alpha carbons of the first and last residues were selected as fixed and pulled atoms, respectively. The system was prepared with the Protein Preparation Wizard tool (Sastry, Adzhigirey et al. 2013), adding missing hydrogen atoms, fixing environment dependent protonation states and checking disulfide bonds. PELE parameters set up for these simulations were the standard set for proteins with OPLSAA (Jorgensen, Maxwell et al. 1996) force field combined with the VDGBNP implicit solvent and 0.15 M of ionic strength. Pulling parameters were 300 kcal/(mol·Å^2) and 10 kcal/(mol·Å^2) for the harmonic force constant of the fixed and pulled atoms, respectively. Pulling speed can not be determined in

MC simulations because time is not simulated. For this reason, we defined the pulling speed as the displacement per accepted step during the simulation updating the virtual bead position just when the movement has been accepted. Thus, we found the optimum speed 0.01 Å/step to reproduce similar MD force-extension profiles. We performed three PELE independent trajectories to reduce the noise and the possible bias associated with the initial structure, and combined these three trajectories in a single one using a binning method.

We performed an MD trajectory of the same initial structure to check PELE's accuracy simulating force-extension profiles using NAMD 2.9 (Phillips, Braun et al. 2005). MD simulations have been performed using explicit solvent with a spherical water box of TIP3P water molecules (Jorgensen, Chandrasekhar et al. 1983) with 50 Å radii. Force field used to parameterise systems topology was c31b1 release of CHARMM (Vanommeslaeghe, Hatcher et al. 2010). The equilibration protocol consists of two minimizations: first, minimize just waters and then a global minimization of the whole system. Then, we performed 200 ps heating up the system to 300 K using a weak-coupling algorithm with constant pressure. Constant pressure and temperature (NPT ensemble) has been applied to the system using a Berendsen barostat and thermostat (Berendsen, Postma et al. 1984) combined with a time step of 2 femtoseconds in the equilibration and production process. Checking the convergence of total energy, temperature and pressure the simulations have been considered equilibrated after one nanosecond. SMD parameters were 7 kcal/(mol·Å^2) for the harmonic force constant with a pulling speed of 2.5 Å/picosecond and the pulling direction was the initial direction between the fixed and the pulled atom.

4.1.1.2 Closure

When comparing the final PELE and SMD force-extension profiles it clearly shows large similarities (see Figure 39). Moreover, the initial minimum corresponding to hydrogen bond breaking of two folding domains, agrees well with the experimental AFM profile (Carrion-Vazquez, Li et al. 2003), and provides an atomic detailed explanation about the origin of these force peaks. MD cannot measure the rest of experimental Ubiquitin peaks due to the large pulling speed (a few orders of magnitude) needed by the computational resources respect to the AFM experiment. As PELE have been calibrated to reproduce a similar behavior as MD at nanosecond scale the result shows the same problem. Last part of the force-extension profiles in Figure 39 (>280 Å) shows an enormous increasing force produced by the covalent bonds when the protein is totally unfolded.

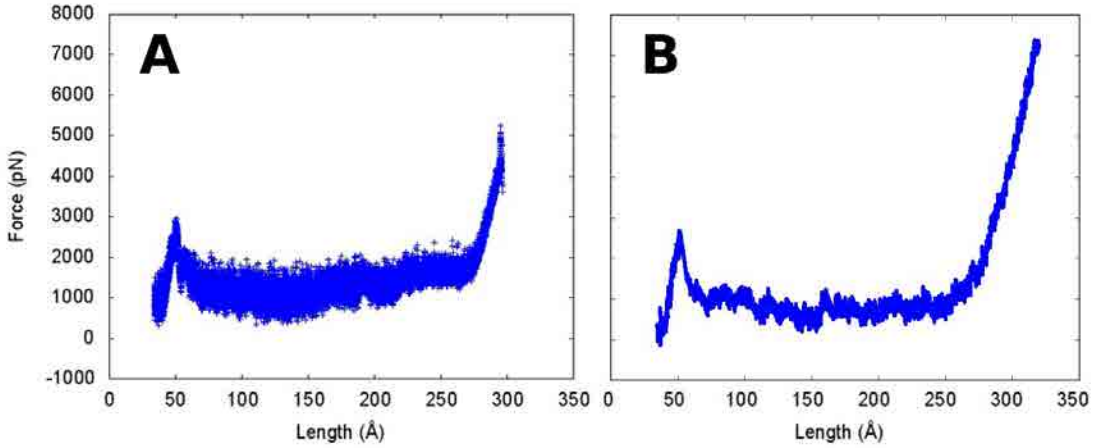


Figure 39. Ubiquitin force-extension profiles. Panel A, average of three PELE independent pulling trajectories. Panel B, SMD trajectory generated with NAMD.

4.1.2 PELE test case II: azurin

Holo- and apo-azurin (with and without coordinated CU, respectively) display nearly identical tertiary structure (Nar, Messerschmidt et al. 1991, Nar, Messerschmidt et al. 1992) and thus provide an opportunity to directly determine the role of the metal in azurin (Az) mechanical stability using AFM. Experiments were performed (by our collaborators) on monomeric Az for several reasons, despite the difficulty of the recordings and data analysis compared to multidomain proteins often used in AFM. Monomers are more biologically relevant and enable a direct comparison with bulk experiments performed with cupredoxins. In addition, using wild type monomeric Az allows avoiding structural alterations introduced by molecular handles, domain-domain interactions and aggregation problems of multidomain proteins. Finally, monomeric Az allowed direct comparison with theoretical simulations.

The variability observed in Az AFM experiments could not be reduced by increasing the number of experiments, probably due to the concurrence of intrinsically variable conditions like the structural configuration of the protein and the different attachment residues to the AFM tip. In order to gain insight into these variables, we turn into molecular simulations by using PELE to obtain Az unfolding curves (Figure 40A).

4.1.2.1 Calculations and discussions

The crystal structure 4AZU (Nar, Messerschmidt et al. 1991) was selected from the protein data bank for the computational simulations. The system was prepared like Ubiquitin in the previous section, with the Protein Preparation Wizard tool (Sastry, Adzhigirey et al. 2013) adding missing hydrogen atoms, fixing environment dependent protonation states and

checking disulfide bonds. PELE used the OPLSAA (Jorgensen, Maxwell et al. 1996) force field with the implicit surface generalized solvent model VDGBNP (Zhu, Shirts et al. 2007). The charge of the Cu ion was set to +2 and the ionic strength to 0.15 mol/dm^3. Due to the qualitative nature of our simulations, the metal coordination bonds were described only by means of the force field electrostatic term. To validate this approach, we performed preliminary tests using a model with the Cu center plus the 5 coordinated ligands, where we built gas phase pulling energy profiles in each coordination bond with both QM(M06/lacvp**) and OPLS levels of theory. Results were actually surprisingly good and in 4 of the ligands (the two His, the Met and the Gly, see Figure 41) the energy profiles were in qualitative (and even semi-quantitative) agreement with quantum calculations; only breaking the Cu-Cys bond (charge separation) resulted in significant off-results (4x larger OPLS dissociation energies). Moreover, these gas phase differences were further reduced when adding electrostatic screening derived from the "condensed" protein media.

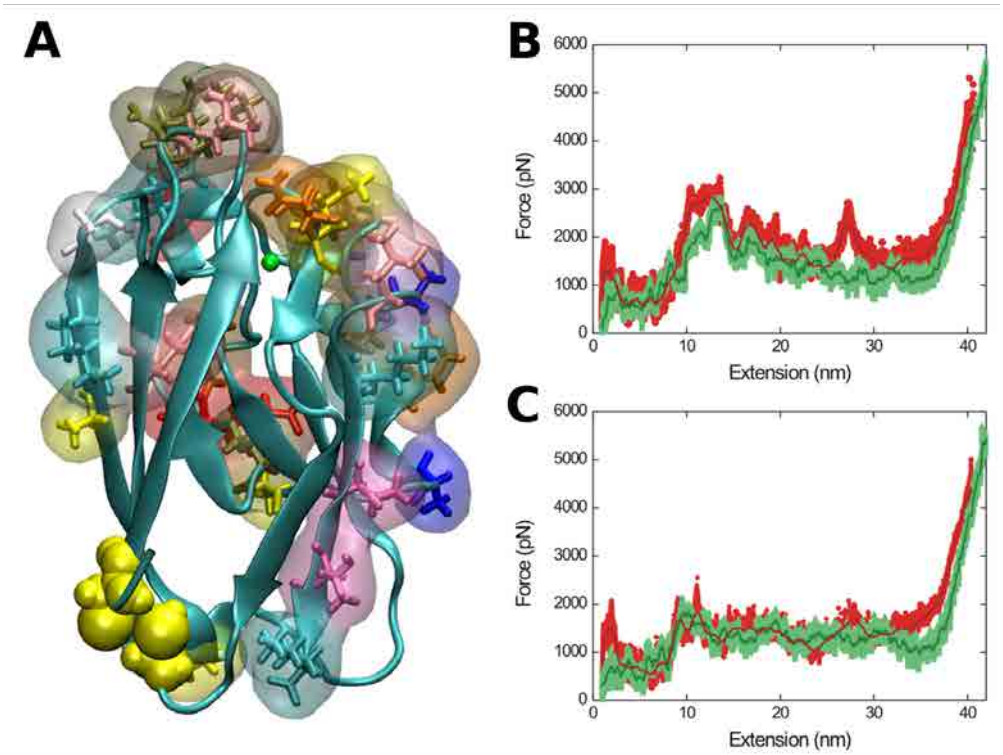


Figure 40. Panel A, 3D structure of Pseudomonas aeruginosa Az where all residues analysed are highlighted. The substrate-attaching residues Cys3 and Cys26 (yellow balls) were fixed in the simulations. The copper ion is represented by a green sphere in the region facing the AFM probe. Colour code for the highlighted residues: Met (yellow), Asp (red), Ala (blue), Asn (dark green), Lys (cyan), Glu (pink), Gln (orange), Gly (white). Panel B and C, comparison between SMD (green) and PELE (red) force-extension profile for the holo (B) and apo (C) states, for the pulling of residue Lys128. The solid lines correspond to the average force fit using 300 points.

As in the previous case, in order to validate our new approach with an established technique, we performed SMD simulations for Lys128. SMD was performed again using NAMD 2.9 (Phillips, Braun et al. 2005) package and a modified version of CHARMM22

(Vanommeslaeghe, Hatcher et al. 2010) force field to include the CU parameters and a spherical water box, with 60 Å radii. System was equilibrated with an initial minimization followed by a heating process increasing every 0.4 ps 10 degrees from 0 K to 310 K. Then, the alpha carbon of Cys26 was fixed and the gamma carbon of Lys128 was pulled at 0.5 Å/ps with a 7 kcal/(mol·Å^2) (486.36 pN/Å) of spring constant. Simulation was carried out using constant pressure and particle mesh Ewald (PME). Total pulling simulation time was 1 ns. Figure 40 B and C shows the comparison between SMD (green) and PELE (red) force-extension profile for the holo and apo states, for the pulling of residue Lys128. As seen in Figure 40 B and C, SMD provides the same results as PELE but at the expense of approximately five times higher CPU cost.

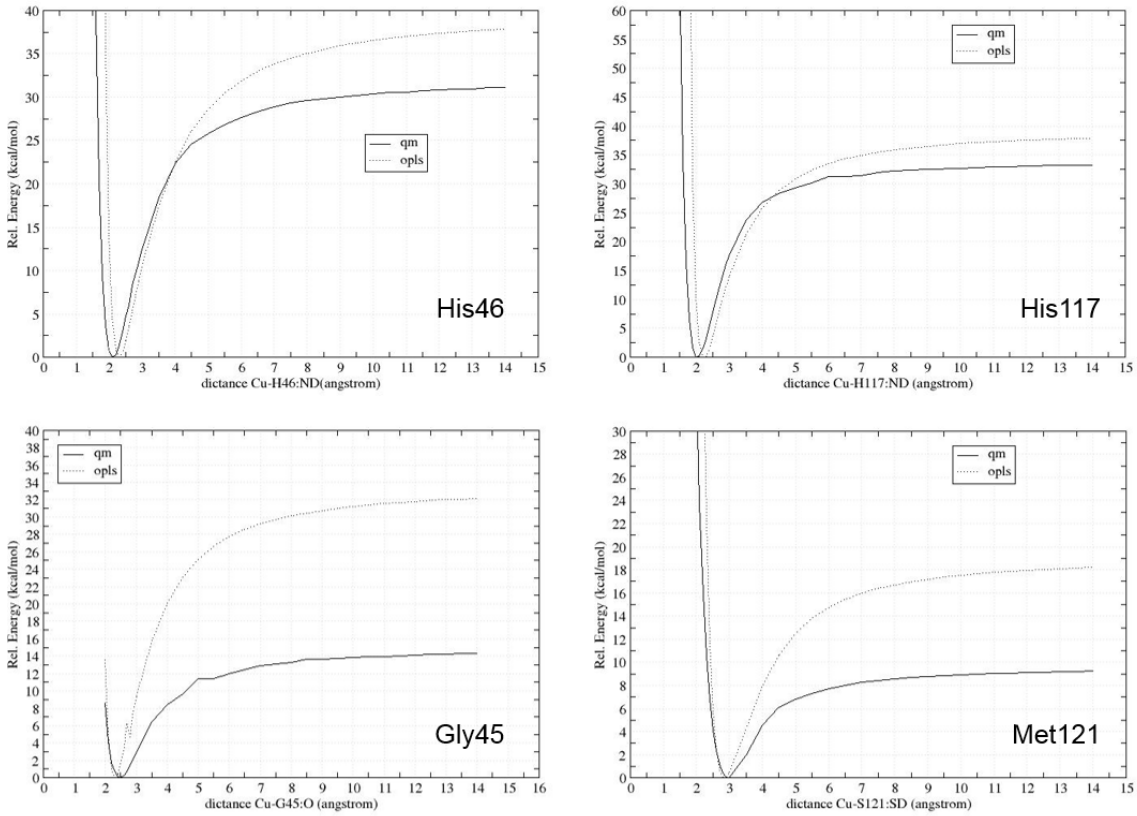


Figure 41. Metal coordination bond analysis. Comparison of QM and OPLS dissociation energy profiles along the Cu-X coordination bond for X: His46, His117, Gly45 and Met121.

Once PELE results were validated with SMD we performed three independent trajectories for each selected pulling residue and state (apo/holo). From experimental AFM, it is not possible to determine which residue is attached to the cantilever. For this reason, each simulation was performed with a selected residue (to be pulled) from a surface list. The surface residues included: Gln12, Met13, Leu33, Asn38, Leu39, Lys41, Asn42, Val43, Ala54, Gln57, Val60, Ala65, Asp69, Pro75, Asp76, Asp77, Ser78, Val80, Gly90, Lys92, Ser94, Ser100, Pro115,

Gly116, Ala119, Leu120, Lys122, Thr124, Thr126, Lys128 (Figure 40A). Additionally, the atom to be restrained was chosen randomly between the carbons of the side chain. Then, the average force with respect to the extension was linearly interpolated in order to obtain a continuum force plot. The force peak corresponding to the largest difference in extension between holo and apo-Az simulations was then selected as an indication of the "rupture force" (Figure 42C). Thus, for each one of the residues analyzed (Figure 40A) there are three pairs of points in Figure 43B (each pair containing one apo and one holo point from each independent simulation). Notice that by using this "rupture force", instead of a fixed rupture force, we obtain possibly an upper bound value for the differences between apo and holo.

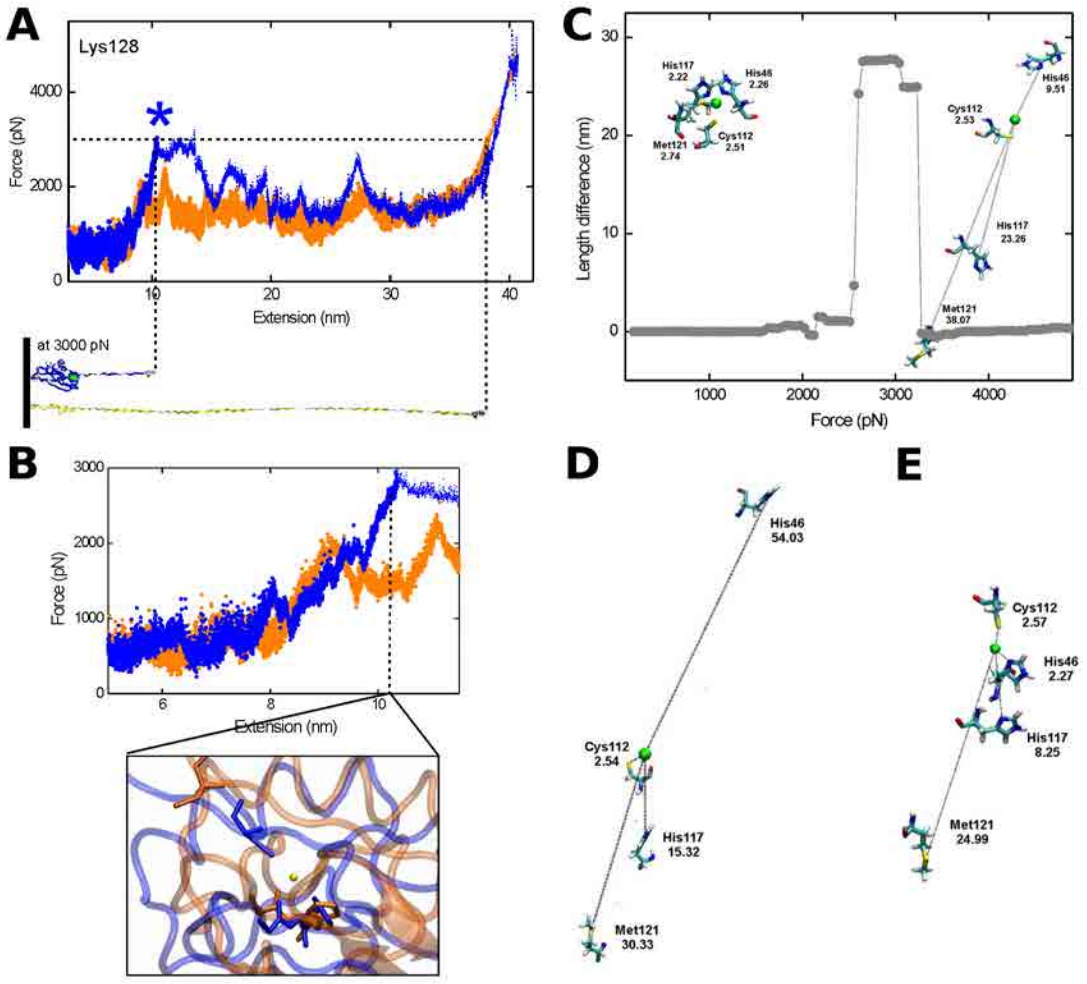


Figure 42. PELE force-extension and molecular view of the force-extension. Panel A, PELE force-extension profile for surface residue Lys128 of holo-Az (blue) and apo-Az (orange). The asterisk (*) indicates the maximum force-extension value obtained for holo-Az. Panel B, initial stage of unfolding and Cu-binding site conformation snap-shots for holo-Az and apo-Az at 10 nm extension for Lys 128. Panel C, difference in extension (length difference) between the holo and apo-Az obtained by PELE at every given force for residue Lys128. Panel D, schematics of the distance between the Cu atom and four of its coordination residues (His46, Cys112, His117 and Met121) before and after the maximum force peak at 3000 pN for pulling residues Lys128 (C), Ala65 (D) and Pro75 (E).

Due to computer limitations (classical force fields, pulling speed, etc.), however, simulations result in force overestimation (Rico, Gonzalez et al. 2013) and AFM modelling is based mostly in qualitative explorations (Lu and Schulten 1999). In Figure 42 A, B we present example profiles obtained for surface residue Lys128 in holo- and apo-Az simulations. These plots clearly show that forces in the holo model are higher than in the apo model for a large fraction of the trajectory. The difference in extension at a constant force is shown in the snapshot (structure of partially unfolded protein) of Figure 42A and is calculated in Figure 42C for the entire range of force. As observed in the holo and apo-Az structure in Figure 42 A, while apo-Az is almost fully extended, only part of the holo-Az is unfolded at the selected force. Figure 42C shows the holo and apo-Az difference in extension length between fixed (Cys26) and pulled (gamma carbon of Lys128) atoms for every given force. This difference is highest at 3000 pN, as a result of a shorter extension in holo-Az due to the Cu interaction with its coordinating residues. Figure 42 also includes two snapshots of the atomic representation showing the metal coordination distances before pulling and after the peak for the pulling residue Lys128 (Figure 42C), and the final snapshot for residues Ala65 (Figure 42D) and Pro75 (Figure 42E), which display a markedly different behavior.

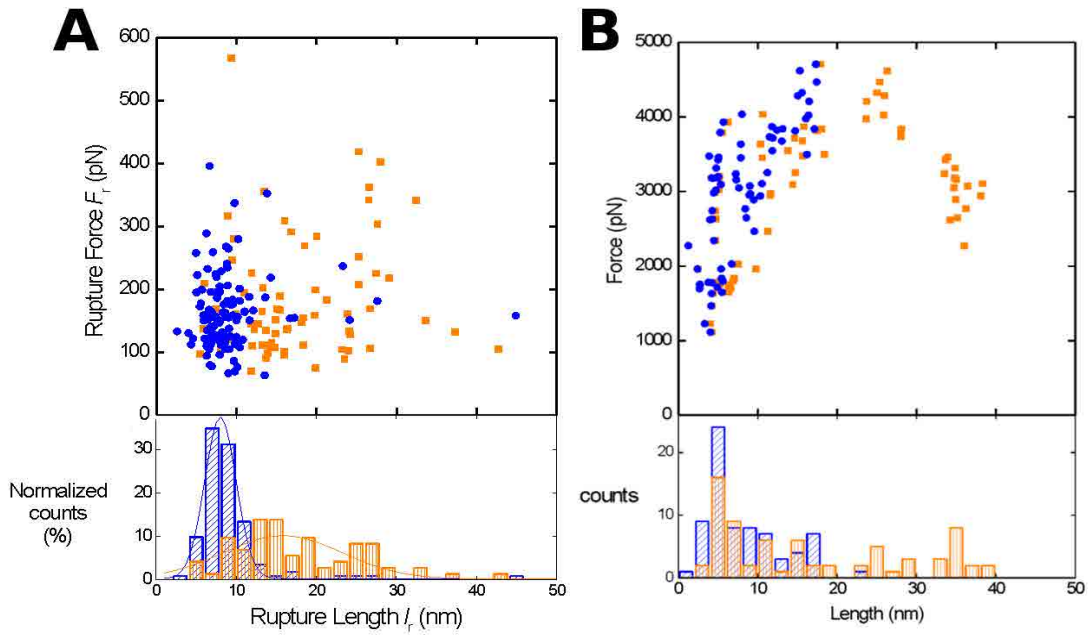


Figure 43. Panel A, experimental distribution of rupture force (Fr) and length (lr) for AFM of individual holo-Az (blue) and apo-Az (orange) in 50 mM ammonium acetate buffer pH 4.5, at 25 ºC. Panel B, force vs. length obtained from all PELE simulations.

The expanded view of Figure 42B shows that during holo-Az unfolding, the force increases abruptly at an extension of 9 nm, whereas apo-Az unfolds at relatively constant force in this range. Simulations were repeated for all residues shown in Figure 40A and the results are summarized in a force versus length plot (Figure 43B) that reproduces the experimental observations of Figure 43A.

4.1.2.2 Closure

Compared to experimental curves, which sample several attachment residues on the protein surface and must be analyzed statistically, in simulations, unfolding events can be individually tracked. Remarkably, both methods fully unfold apo-Az up to 40 nm, whereas holo-Az unfolding is restricted to (or the simulated force increases steeply at) lengths below 10 nm. The shorter extension in holo-Az is due to the Cu interaction with its coordination residues. For every simulated attachment site, the divergence between the apo- and holo-Az extension is accompanied by strain and eventual rupture of metal coordination bonds in the holo-case (Figure 42B, Figure 42C), following different unfolding sequences (as exemplified in Figure 42 C, D, E for pulling residues Lys128, Pro75 and Ala65). Together, these results indicate that the metal binding region is mechanically flexible when the metal is not coordinated, and Cu coordination prevents the full extension of the protein regardless of the attachment site. Our results are in accordance with reported observations of conformational heterogeneities for the metal binding site in cupredoxins in absence of the metal and their suggestion of the contribution of the metal ion to the rigidity (Messerschmidt, Prade et al. 1998, Ryde, Olsson et al. 2000, Zaballa, Abriata et al. 2012, Abriata, Vila et al. 2014). Indeed, several copper-mediated protein-protein interactions have been identified in recent years (Fu, Tsui et al. 2013, Girotto, Cendron et al. 2014) and in some cases the crystal structure involve the copper ion along with direct interactions between large protein surfaces (Banci, Bertini et al. 2006). Furthermore, a deformation of the metal binding site upon metallochaperone binding was observed by NMR (Abriata, Banci et al. 2008) and is in agreement with MD simulations showing great flexibility in apoAz and especially in the binuclear CuA domain of cytochrome c oxidase (Abriata, Vila et al. 2014).

4.2 MCPRO to stretch molecules

MCPRO (Monte Carlo for proteins) (Jorgensen and Tirado–Rives 2005) is an algorithm developed for proteins using BOSS (Jorgensen and Tirado–Rives 2005) as a frame. BOSS program (an acronym for Biochemical and Organic Simulation System) is a general purpose molecular modelling for molecular mechanics simulations developed for small systems.

MCPRO applies the BOSS molecular mechanics functions to work with residues performing an efficient energy calculation of large systems such as proteins or DNA. MCPRO simulations are carried out at any specified temperature and at constant pressure (NPT) or at constant volume (NVT) ensemble. Evaluation of total energy of the system is performed using the OPLS-AA force field (Jorgensen, Maxwell et al. 1996). Solvent contribution can be explicit using different solvent models such as TIP3P, TIP4P (Jorgensen, Chandrasekhar et al. 1983) and TIP5P (Mahoney and Jorgensen 2001) or approximated by a GBSA (Qiu, Shenkin et al. 1997) implicit solvent. MCPRO uses a solvent sphere (caps) to reduce the number of water molecules in explicit solvent simulations and increase the speed up for protein-ligand complexes. Solvent molecules are placed on the system using precomputed symmetric boxes to reduce the equilibration steps for the simulation. MCPRO uses internal coordinates to represent the molecules in the zmatrix format to perturb the ligand and protein structure.

At each MCPRO step, one random residue of the system is selected to be perturbed. If the residue is a solvent molecule, one random translation and rotation is applied. Translation and rotation size is determined during the equilibration procedure to ensure a 40% of acceptance in the sampling. Moreover, preferential sampling (Owicki and Scheraga 1977) is applied to increase the sampling of the solvent molecules around the compound. If the residue selected belongs to the protein, the bonds, angles and dihedrals of the side chains are perturbed while the backbone is held fixed. Backbone of proteins (or DNA) is perturbed using the concerted rotations procedure (CRA) (Ulmschneider and Jorgensen 2003) with a specified frequency. Moreover, MCPRO also performs random translations and rotations of the ligand and protein separately using a predefined set of atoms to define the centre of the rotations. It allows the ligand exploration of different conformations in the binding site. Each new step is accepted or rejected following a metropolis criteria based on the total energy difference between the system and the specified temperature.

Atom pulling of molecules during the MCPRO simulation is carried out in the same way than PELE adding a harmonic potential to the total energy where two atoms are specified by the user as fixed and pulled atoms. The direction of the pulling is determined as the line between these two atoms and can be computed once at the beginning of the simulation or can be updated in each MCPRO step (accepted or rejected). In the initial step of the simulation, one virtual bead is generated in the position of the pulled atom, and a distance harmonic constraint is generated between them with a zero equilibrium length. Each accepted step increases, in the pulling direction, the distance between the fixed atom and the virtual bead by a constant displacement (Δr). The energy of the harmonic constraint bias the MCPRO acceptance to displace the pulled atom following the pulling direction. The contribution of the force in the

pulling direction is computed as the scalar product between the unitary direction vector and the harmonic force vector (see pulling scheme in Figure 38).

The addition of a dynamic harmonic constraint term to the final energy function produces a decrease in the acceptance. For this reason, the force constant value and the pulling speed must be selected according to the system. A high value of the force constant produces a strong bias to the pulling direction producing in some case artifacts in the structures. Hence, large values of Δr produce non-realistic movements of the atoms producing jumps of the atoms in the pulled direction. These two problems overestimate the value of the average force during the pulling process.

Absolute binding free energy

MCPRO simulations have been successfully combined with free energy perturbations (FEP) technique to evaluate relative binding free energies. In this work, we have combined MCPRO simulations with the pulling of atoms in a similar scheme to SMD. We aim to connect the non-equilibrium work produced during the pulling process with the equilibrium free energy between two states with MCPRO. To this aim we will use Jarzynski's equality (Jarzynski 1997) over a set of independent MCPRO trajectories to study the PMF corresponding to the reaction coordinate defined by the distance between the two extremes of the chain.

Pulling force generated in each MCPRO step is computed by $\vec{F} = k \cdot (\vec{r} - \overrightarrow{r_0})$ where $\vec{r} - \overrightarrow{r_0}$ corresponds to the distance between the virtual bead and the pulled atom. Force contribution into the reaction coordinate is computed due to the scalar product between the force vector and one unitary vector in the pulling direction. Work is computed at each step of the reaction coordinate as $W_i = \overrightarrow{F_i} \cdot \Delta\vec{r}$ where $\Delta\vec{r}$ is the constant displacement applied in each MCPRO accepted step. Total work of a trajectory is computed as the sum of the works generated in each step ($W = \sum_{i=0}^{N} W_i$).

The free energy (ΔF) comes from the exponential average work by Jarzynski as mentioned in chapter 3. The exponential average has been compared with the average work and dissipated work ($W_{diss} = W_{avg} - \Delta F$). For a set of simulations, it is essential to have a dissipated work as low as possible. Pulling speed, expressed as displacement step in MCPRO, and force constant are two variable parameters. Large force constant will produce larger fluctuations (Park, Khalili-Araghi et al. 2003) but the results, if converged, should be similar. In any case, it is recommended to select large force to ensure small deviation between the pulled atom and

the virtual bead but not much bigger than that. The choice of the pulling speed depends on the system but must be selected slow enough that a small set of trajectories is sufficient for converge; a fast pulling speed will increase the distance of the equilibrium, and more trajectories will be required to converge the result.

4.2.1 MCPRO test case: unfolding free energy of deca-alanine

Helix-coil transition in deca-alanine is a classical example used to test MD methods such as umbrella sampling (Torrie and Valleau 1977) or SMD (Park, Khalili-Araghi et al. 2003). It is a simple alpha helix composed of ten alanine residues small enough to simulate many trajectories in an affordable time and complex enough to be considered a prototype of a biomolecule. Relaxation time is short and allows the study of the helix-coil transition in a reversible way.

Deca-alanine molecule was generated from the sequence with PEPz (Jorgensen and Tirado–Rives 2005) application in combination with a PDB folded structure generated using Maestro (2015). MCPRO performs the fluctuations in internal coordinates using the backbone nitrogen of the fourth residue as the center for the perturbations. For this reason, we have constructed a deca-alanine with thirteen residues, and ignored the movement of the first three residues. Thus, our fixed atom corresponds to the nitrogen in the fourth residue, and the pulled atom is the backbone atom corresponding to the O with PDB name O2 of the last residue. Simulations were performed in vacuum where deca-alanine shows a stable helix conformation.

We used the same optimum force constant 7.2 kcal/(mol·Å^2) found in a previous study of deca-alanine unfolding using SMD. To find the optimum MC pulling speed (displacement/accepted step), we tried a set of different values to analyze the convergence of the work-extension profile. Figure 44A shows how the work associated with each speed for the deca-alanine system tends to converge to a value around 25 kcal/mol when the pulling speed is reduced. Using the lowest speed in an affordable computational time (10^{-5} Å/step), we performed 60 independent simulations starting from the same initial structure to evaluate the unfolding free energy from these non-equilibrium trajectories with Jarsynski's equality. As our simulation started at 15 Å of end-to-end atom separation and the published result started at 13 Å result was shifted down 2.5 kcal/mol to add this initial work contribution to our findings. Final unfolding free energy evaluated with MCPRO was 26.5 kcal/mol showing a difference of 5 kcal/mol with the reference value extracted from the reversible pulling with MD (21.5 kcal/mol) (Park, Khalili-Araghi et al. 2003). As it is expected, MCPRO shows a

slightly higher value of work than the MD reversible work for the whole trajectory produced by the pulling speed and the conformational sampling.

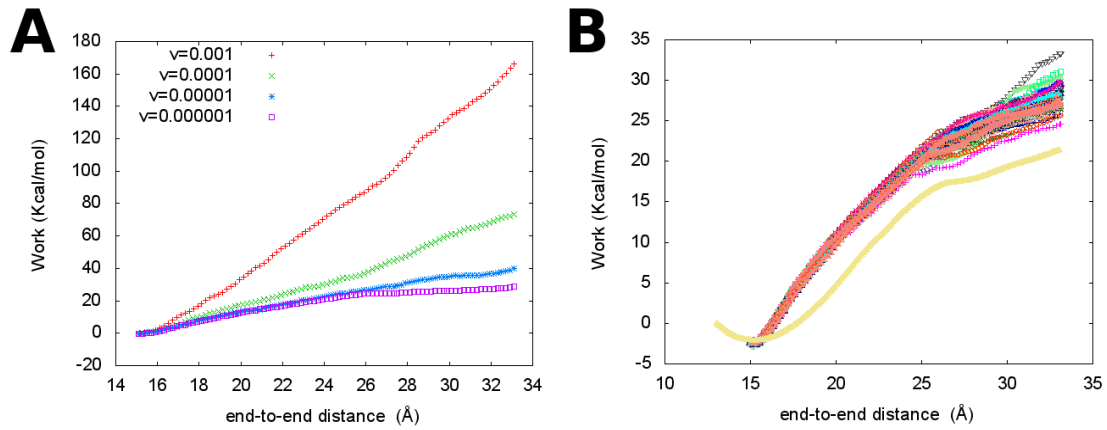


Figure 44. Deca-alanine unfolding work computed with MCPRO. Panel A, unfolding work computed for different pulling speeds. Panle B, 60 independent MCPRO unfolding work trajectories computed for the optimum pulling speed (10^{-5} Å/accepted step). Yellow line corresponds to the PMF associated to reversible pulling computed with SMD.

Overall, our results indicate how MCPRO is capable of computing unfolding free energies similar to MD, opening the door to free energy studies based on steered MC approaches. Future arrangements will be the application of MCPRO to evaluate global binding free energies for protein/DNA-ligands though the pulling out of the binding site.

Chapter 5

Multiscale approach for protein-protein interactions: CG sampling and all-atom refinement

Structural prediction of protein–protein complexes given the structures of the two interacting compounds in their unbound state is a key problem in biophysics. In addition to the issue of near native orientations sampling, one of the modelling main difficulties is to discriminate true from false positives. We aim to expand the protein–ligand interaction study and to test the goodness of an all-atom force field when scoring protein–protein docking poses. To this end, we have developed a multiscale protocol based on a CG approach to speed up the protein-protein sampling combined with an all-atom optimization to improve pose discrimination. For this reason, we applied the CG model published by Basdevant et al. (Basdevant, Borgis et al. 2007) where each residue is represented by one, two or three beads depending on the residue size. Then, we developed an MC sampling algorithm based on discrete random jumps combined with an acceptance criteria based on residue distances and CG energies (see details below). Later, we transformed the best CG poses (in terms of CG energy) to all-atom structures to perform another filtering based on the all-atom binding energy of the optimized structures.

5.1 Multiscale protocol

5.1.1 CG sampling

CG models have been developed to speed up expensive computational algorithms, such as MD or protein-protein docking. These models aim to reduce the degrees of freedom for the sampling keeping the essential behavior of the original mathematical model. Basdevant CG model only takes into account non-bonding interactions using two terms to describe the interaction (VDW and Coulomb). Thus, CG sampling is performed with rigid body exploration. CG algorithm is divided into three main steps: (i) initialization of the system, (ii) global search and (iii) local search (see Figure 45). In initialization, CG code generates beads from a PDB structure, determines the beads of the surface and computes maximum restraint distance allowed (R^{max}) between some selected residues (restraint distance criteria). Global

search step explores a large part of the protein surface trying to identify the binding site and each step consists of big random translations (up to 30 Å) and rotations (up to 360º). A global search step is accepted if the move reduces the distance defined in the restraint distance criteria (R). CG algorithm iterates global search steps until the conformation accomplishes the restraint distance criteria. Then, the local search algorithm is used mixing big translations (typically around 4-5 Å) with small ones (up to 3 Å) and rotation (up to 10º) to explore faster the binding region found and avoid over exploration of local minima. Acceptance or rejection of each new step is based on three main criteria: restraint distance criteria, (beads) overlapping criteria and CG binding energy (see below). These three conditions (called filtering in the scheme) reduce search space avoiding clashes and solutions that are far from the binding site. In a few hours, CG sampling can generate more than 100.000 accepted conformations.

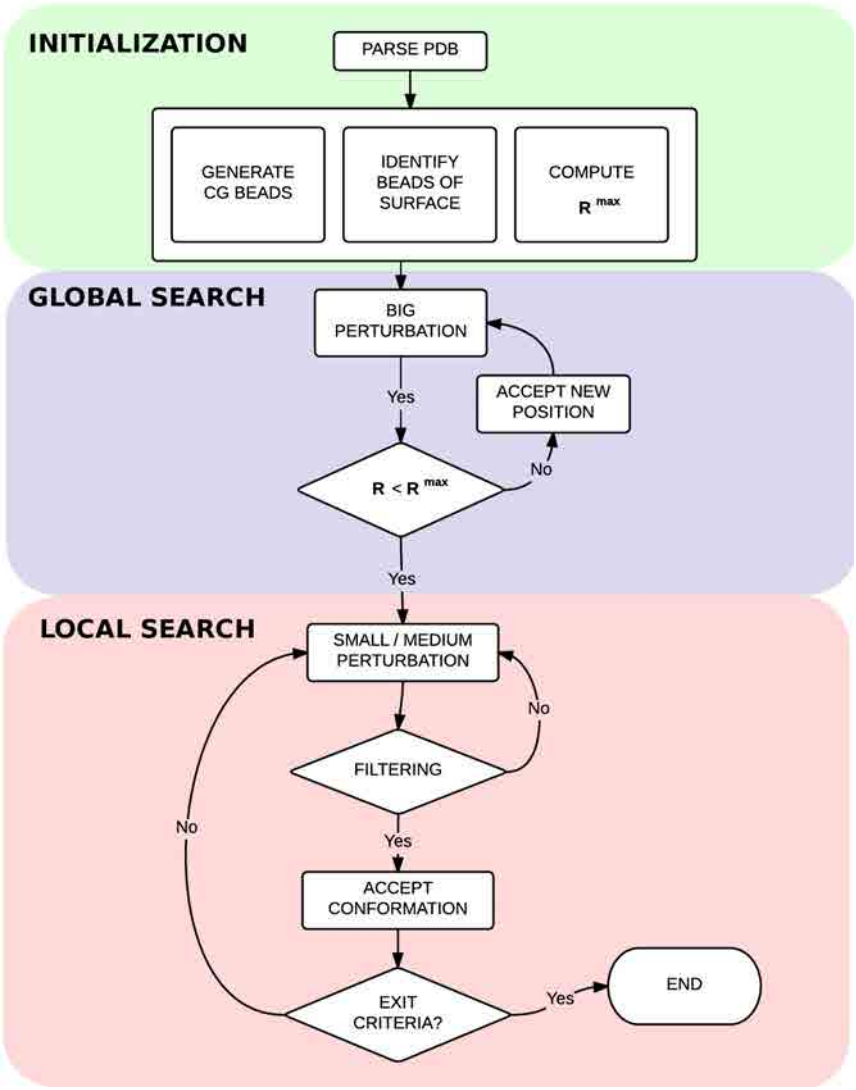


Figure 45. CG sampling algorithm scheme.

Beads generation

As mentioned before, Basdevant CG model (Basdevant, Borgis et al. 2007) reduces all the atoms of a residue to two or three beads depending on the residue size. Each bead is composed of three Van der Waals parameters and one charge parameter. VDW terms use the energy parameter (ϵ_{ij}), a length parameter for repulsion (λ_{ij}) and a Gaussian parameters term for attraction (σ_{ij}) (see equation 10). Standard residue parameters have been extracted from Basdevant et al 2007. Cofactors parameters have been generated using an average distance between the geometric center and the atoms belonging to the bead. Choice of atoms included in each bead was performed by visual inspection (See Figure 46).

$$V_{vdw}(r_{ij}) = \epsilon_{ij}\left[\left(\frac{\lambda_{ij}}{r_{ij}}\right)^6 - e^{-\left(\frac{r_{ij}}{\sigma_{ij}}\right)^2}\right] \quad (10)$$

The charge of all beads has been recomputed using the AMBER03 (Duan, Wu et al. 2003) force field parameters to get a balanced description of the interactions between proteins and cofactors. We also tried OPLS-AA (Jorgensen, Maxwell et al. 1996) charges but it produced a lot of neutral beads in the residues due to the nature of the charge distribution of this force field. Bead charge was estimated using the average charge of the atoms involved in the bead representation. Because of this, final beads keep total net charge of each residue. The energy pair calculation was performed using a 12 Å cutoff. Beads pair list is generated and updated during the sampling including all beads of a residue if one bead of the residue is inside the cutoff to avoid charge polarization. Parameters fitted for CG model provides an approximation of the energy in the vacuum. To take into account the solvent interaction we have implemented an implicit solvent model called distance dependent dielectric. Moreover, we tried different values of the permittivity for surface and buried beads.

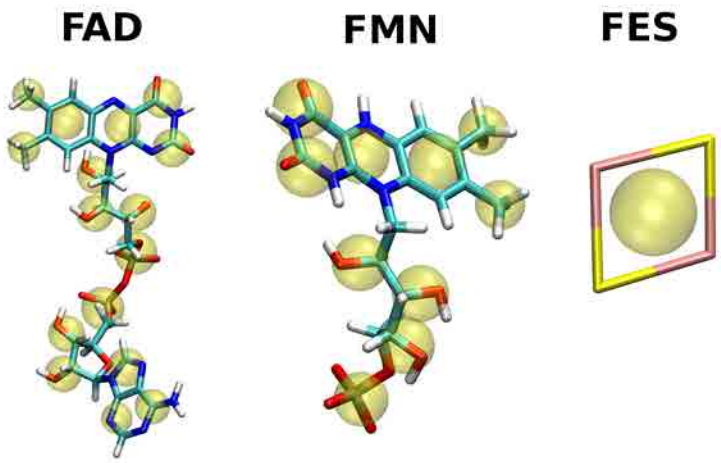


Figure 46. Beads representation of the FAD, FMN and FES cofactors on FNR, Fld and Fd, respectively.

Bead clash and overlapping criteria

An overlapping term has been introduced to reduce geometric error contribution during the sampling due to the CG approximation. This term is computed by Equation 11, allowing us to accept a certain level of manually penetration between beads during clash detection. If distance between beads (D) is bigger than the scaled sum of beads radii (R_{vdwA}, R_{vdwB}) the structures will not be rejected. The optimal parameter found for our test cases for this term was 0.8, which means penetration is allowed by 20 % between beads.

$$D = (R_{vdwA} + R_{vdwB}) \cdot 0.8 \tag{11}$$

CG binding energy scoring

Equation 12 corresponds to the binding energy where E_{AB} corresponds to the total energy of the complex, E_A the isolated monomer A and E_B the isolated monomer B including solvent interactions, respectively. This formula has been used for CG and all atom calculations to accept, reject or score poses using CG and all atom energies, respectively.

$$E_{Bind} = E_{AB} - (E_A - E_B) \tag{12}$$

5.1.2 All-atom refinement

All atom scoring based on OPLS force field can not be computed directly with the structures generated by CG sampling due to possible atomic clashes, wrong protonation states, broken hydrogen bonds. To fix these problems, Protein Preparation Wizard (Sastry, Adzhigirey et al. 2013) was used to protonate using PROPKA algorithm (Olsson, Søndergaard et al. 2011) and to optimize hydrogen bond network interactions of the complexes. This all-atom refinement has been tested improving scoring functions in protein-protein docking (Masone, Cabeza de Vaca et al. 2012) (see 5.3 section). Structures were minimized and scored using PELE (Borrelli, Vitalis et al. 2005, Madadkar-Sobhani and Guallar 2013) to find the best energy of each configuration but applying the OPT protocol instead of PELE protocol (see 5.3 section for details).

5.2 Validation of the All-atom refinement

In order to validate our all-atom refinement protocol we will show that simple algorithms for hydrogen bonds optimization, followed by energy minimization, are enough to rescore a 10 Å RMSD near native pose in 70 % of the cases. With further side chain sampling and taking into account backbone responses to perturbations, we will show that for even difficult cases the near native conformation is identified as the lowest energy one. At the end, we selected the optimal method in terms of accuracy versus performance to include in our multiscale approach.

5.2.1 Calculations and discussions

Pose generation

We used as starting point a set of complexes obtained with FTDock (Gabb, Jackson et al. 1997) and pyDock (Cheng, Blundell et al. 2007), which combined form a well-established protein–protein docking pose generator and scoring function, to generate 10.000 poses for the 84 cases with known X-ray structures both for the unbound and the bound subunits of the protein docking benchmark 2.0 (Mintseris, Wiehe et al. 2005) (as well as for the 176 ones present in benchmark 4.0 (Hwang, Vreven et al. 2010), see below), sets commonly used to evaluate docking algorithms. We selected those cases that included at least one near-native solution within the top 100 poses as ranked by pyDock. On the basis of our previous knowledge of protein–ligand interactions, successful scoring was only possible when low RMSD near-native structures were achieved. Thus, a near-native solution was defined here as that with RMSD less than 5.5 Å for the ligand Ca atoms with respect to the complex structure, after superimposing the corresponding receptor molecules, a criteria loose enough to keep in our study set 12 systems from Benchmark 4.0. Note that this definition is stricter than the ‘‘acceptable’’ criteria in CAPRI (10 Å), and closer to the ‘‘medium’’ criterion (5 Å). This is the list of cases that satisfy the above criteria in Benchmark 4.0: 1b6c, 1buh, 1e6e, 1fsk, 1ppe, 1nca, 1tmq, 1udi, 2sni, and 7cei.

Docking refinement

Three different protocols, with increasing complexity, were applied to refine the docked poses. Prior to the refinement, we checked for possible differences between bound and

unbound proteins, like missing side chains, ions, and added them to the rigid-body orientations generated by FTDock. Furthermore, the poses were visually inspected for unusual structures. The refinement techniques are based on all-atom force field energy interaction, being very sensitive to atomic steric clashes and electrostatic interactions. Interestingly, large electrostatic stabilization might overcome small steric clashes and introduce a false positive. In these cases, clearly wrong structures with large protein–protein penetration or even interface knot-like loops might produce excellent scores. This point is further discussed below.

- *MIN*. The initial protocol is based on a minimization with P.L.O.P. (Jacobson, Friesner et al. 2002, Jacobson, Pincus et al. 2004) using a truncated Newton algorithm with a root mean square gradient (RMSG) of 0.01, an OPLS-AA force field and a surface generalized Born (SGB) implicit solvent. Geometry optimization allows relaxing bad steric contacts but does not add enough sampling to optimize for a hydrogen bond network.

- *OPT*. The second protocol was to use the Schrodinger's Protein Preparation Wizard (Sastry, Adzhigirey et al. 2013) to optimize the overall complex hydrogen bond network. The algorithm first analyzes the system and builds hydrogen-bonded clusters. Two hydrogen bonds are included in the same cluster if their heavy atoms are within 4.0 Å. Then the highest degree of sampling was used, which performs 10^5 Monte Carlo moves for each cluster. The optimization is performed by reorienting hydroxyl and thiol groups, water molecules, amide groups of Asn and Gln, and the imidazole ring in His; and predicting protonation states of His, Asp, and Glu. Each possibility is scored, determining the quality of the hydrogen bond network of the species in the cluster as well as with the surrounding environment. The scoring function is based on simple electrostatic and geometry considerations. The core of the scoring function involves the number of hydrogen bonds and their quality (based on their geometries relative to an idealized hydrogen bond). Additionally, assignments placing two polar hydrogen atoms within 2.0 Å of one another are given a high penalty. After the hydrogen bond optimization, each pose is minimized using the same parameters as in the first protocol. Additionally, due to the approximate nature of the SGB solvent that overestimates ionic contacts (Yu, Jacobson et al. 2004), different ionic strength constants were tested.

- *PELE*. We included the lowest 20 modes, computed including the full complex in the ANM network, from which we selected one at each iteration randomly. The largest

alpha carbon displacement is set up to 0.8 Å (corresponding with the largest alpha carbon coefficient in the mode). This does not mean that the atom moves 0.8 Å, since the optimization ends when a RMSG of 0.01 is reached. Typical backbone RMSD of 0.1–0.3 are achieved in each iteration. Side chain sampling proceeds by placing all side chains local (within 4 Å) to the protein–protein interface with a rotamer library side chain optimization at a rotamer resolution of 10º. The side chain algorithm uses steric filtering and a clustering method to reduce the number of rotamers to be minimized. Minimization, which is the last step, involves the free minimization of the entire system using (again) a RMSG of 0.01, an OPLS-AA force field and a surface generalized Born (SGB) implicit solvent.

For the optimized structures in any of the three refinement procedures, the interaction energies: $E = E_{AB} - (E_A + E_B)$ were calculated, and all poses rescored. E_{AB} is the total complex energy while E_A and E_B are the energies for its respective isolated monomers.

Results and discussion

Figure 47 shows pyDock's scores and the interaction energies obtained for the first two refinement protocols, MIN and OPT, applied to the top 100 selected poses for the 2SNI, 1UDI, and 1NCA systems. The plots for the remainder of the systems are shown in Masone et al (Masone, Cabeza de Vaca et al. 2012). The energies for the reference complex, after applying the same refinement procedure, are also indicated with a larger square symbol. All RMSD, however, are shown to the non-minimized reference crystal structure. Thus, the refined reference complex deviates slightly from the zero of RMSD. While pyDock is capable of scoring well some low RMSD poses (see also Table 3), a clear improvement in the correlation RMSD vs. interaction energy/score is observed, when comparing the first two refinement methods with pyDock in almost every system. As seen, from the first two plots on the left in Figure 47, the scoring correlation slightly improves when minimizing the complexes, MIN protocol. In particular, some improvement is observed in the region close (within 20 Å) to the native reference, the region where we would expect a correlation between the binding energy and the RMSD. This improvement is particularly seen in the 1UDI and 1NCA systems; the already good correlation of pyDock in the 2SNI complex is well reproduced. The MIN procedure involves approximately 2–5 CPU minutes per pose (depending on the size of the system), but obviously the procedure can be fully parallelized. The most remarkable improvement, however, is obtained when using the hydrogen bond optimization protocol, OPT. As seen when comparing the two right panels in Figure 47, the OPT procedure eliminates several false positives, 2SNI and 1UDI, as well as stabilizes some

near-native complexes, 1UDI and 1NCA. In 8 of the 10 systems, the OPLS-SGB all atom potential is capable of identifying the near-native structure as the lowest energy pose (9 if we take into account the top 5 poses), see Table 3. Thus, the hydrogen bond contact optimization appears to be crucial in order to rescore the poses with an all-atom force field. The cost associated with the optimization wizard is, on average, another 2–5 min per pose, for a total of approximately 5–10 min for a complete structure refinement.

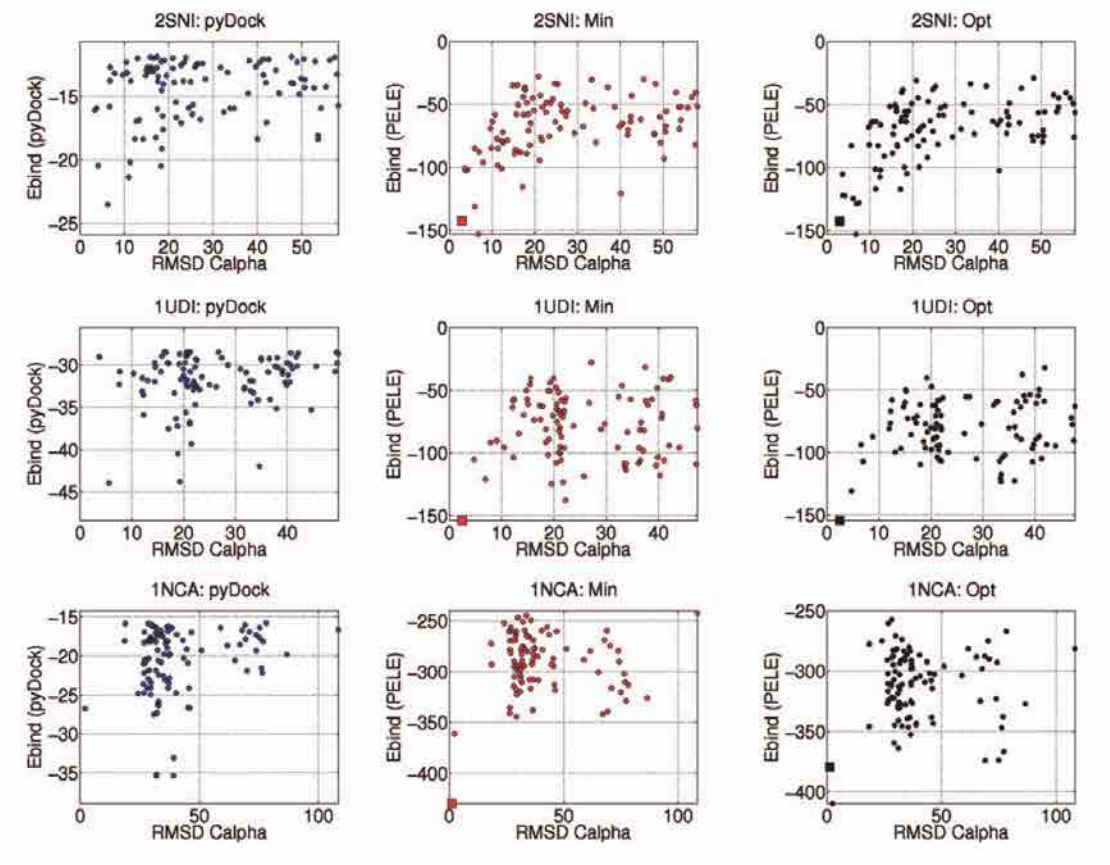


Figure 47. Scoring of three protein–protein complexes. Scoring of 100 poses for three selected protein–protein complexes. Left panel: pyDock, central panel: MIN refinement, and right panel: OPT refinement with 0.45 M ionic strength (see Figure 1 Appendix A for the other systems).

The scoring of the poses in the 1UDI and 1NCA complexes, Figure 47, is a clear example of the importance of the optimization in the interface. In one site, the lowest RMSD poses decrease its interaction energy improving their scoring as a result of the OPT procedure. Additionally, some false positives increase their interaction energy. The mechanism behind both opposite processes, however, is the same: hydrogen bond optimization. Figure 48 shows a detailed view of the flipping of a tyrosine side chain, Tyr32, when optimizing the 1NCA lowest RMSD pose. As seen when comparing the MIN(green), OPT(brown), and reference(blue) structures, the OPT procedure optimizes the alcohol group in the tyrosine and recovers a native hydrogen bond. On the other hand, a false positive in 1UDI (RMSD = 22.21 Å) increases the interaction energy from E_{bind}(MIN) = -137.58 kcal/mol, to score at

E_{bind}(OPT) = -106.71 kcal/mol. We should emphasize here that the OPT protocol allows for hydrogen bonds optimization both internally in each monomer and at the interface. As a result, intrachain hydrogen bond optimization might eliminate interchain hydrogen bonds; the interaction energy increases but the total energy decreases rescoring the false positive.

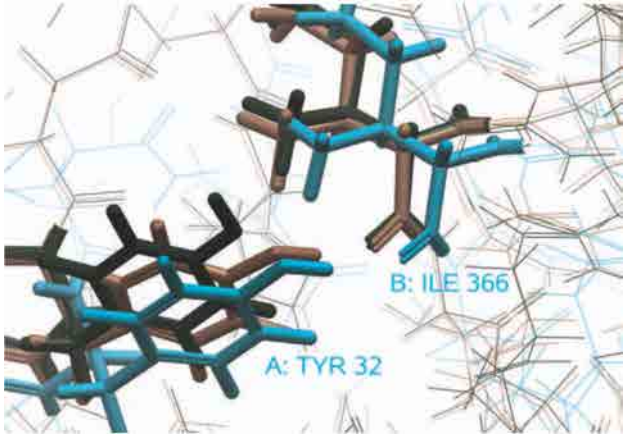


Figure 48. Optimization protocol conformational changes. 1NCA conformational change and h-bond introduced by the OPT procedure. Before OPT (green), after OPT (brown) and crystal (blue).

In Table 3 we summarize the best true positive pose (and in parenthesis, those solutions within 10 Å of the complex) and its ranking for pyDock and the OPT refinement method. The OPT method was run at different ionic strength constant values, showing here those at 0.00 and 0.45 M. PyDock is able to score a native conformation as the top pose in only one case, 1TMQ. In other two cases, the top pose has RMSD < 10 Å. Even in these cases, the energy(score)/RMSD overall correlation is significantly better with the refinement techniques. The OPT refinement, at any ionic strength, significantly improves the top ranking ratio. A large ionic strength (0.45 M), however, slightly improves the scoring for those systems having large charged surfaces, due to the screening of false positives resulting from the overestimation of salt bridges by the SGB implicit solvent method.

Complex name	PyDock's RMSD (Å)	Rank	No ionic RMSD (Å)	Rank	Ionic 0.45 RMSD [Å]	Rank
1PPE	2.5 (9.1)	59 (6)	3.1	1	1.7	1
1UDI	3.7 (5.6)	84 (1)	4.7	2	4.7	1
2SNI	4.1 (6.2)	4 (1)	3.6 (6.8)	3 (1)	4.2 (6.8)	5 (1)
7CEI	5.2	25	5.0 (6.6)	4 (1)	5.0 (6.6)	2 (1)
1FSK	2.8	3	3.7	1	3.9	1
1TMQ	3.5	1	2.9	1	2.9	1
1NCA	2.0	7	2.4	1	2.4	1
1B6C	4.6	3	4.6 (9.5)	5 (3)	4.6 (7.7)	4 (1)
1E6E	3.5	4	3.8 (6.2)	9 (4)	5.4 (6.2)	30 (5)
1BUH	5.2	67	Nan (6.5)	Nan (40)	Nan (6.5)	Nan (44)

Table 3. Benchmark 4.0 pyDock's ranking and ionic strengths evaluation. PyDock's ranking and OPT scoring for different ionic strengths SGB calculations. Solutions are within 5.5 Å RMSD to the native complex, the 10 Å ones and their scoring are indicated between parenthesis.

There are two systems where the refinement method fails to give a top pose within a 10 Å RMSD from the reference crystal: 1BUH and 1E6E. In 1BUH, the quality of the complexes obtained by filtering the top 100 pyDock poses out of 10.000 initial compounds is not satisfactory. We find only three poses under 30 Å and only one under 10 Å. For this system, all scoring functions agree significantly pointing to the lack of suitable candidates. Furthermore, this is a challenging system since both the monomer and complex crystal are missing important residues (THR39-THR47) that when added might generate a significant amount of false positives. For these kinds of systems, a more specific treatment with robust loop prediction techniques and possibly MD loop refinement might be necessary prior to rigid docking.

The 1E6E system required a particular treatment as it includes in its native conformation two hetero groups (iron-sulfur cluster 2Fe-2S and flavin-adenine dinucleotide FAD). The protonation state of the four cluster binding cysteine residues (CYS46 CYS52 CYS55 and CYS92) around the 2Fe-2S hetero group was manually assigned to -1, to give a total charge of 23. The combination of 2Fe-2S with the four cysteines residues were frozen during the minimization process. Again, besides correcting the hetero groups and charges around them, optimizing the hydrogen bond network improved the results significantly. The OPT procedure, however, only scored the near-native pose as the fifth top structure. Thus, we proceeded with a more exhaustive sampling refinement protocol, by using PELE.

PELE refinement

The PELE refinement approach uses a Metropolis MC method were backbone motion and side chain sampling is introduced. In Figure 49 we show the results for the pyDock, MIN, OPT and PELE methods for the 1E6E system. A clear improvement in the correlation as we advance in the refinement method is observed, being the PELE refinement able to score the true native as the top pose, see Figure 49C (green dots). Totally, 500 iterations of PELE were applied to the top OPT 10 poses (plus the reference), with a cost of 8 CPU h per pose in a 2.33 GHz Xenon processor. Side chain sampling involved residues within a 3 Å vicinity of the protein–protein interface. As before, the group of 2Fe-2S with the four cysteines residues was frozen during the PELE sampling. Along the PELE procedure, due to its Metropolis algorithm, there is a deeper optimization of the total energy, which drives also the lowering of the interaction energy as well. All top 10 poses lower their binding energy, but a larger decrease is observed for those poses closer to the reference structure, resulting in a good correlation between the binding energy and the RMSD. In Figure 50 detail of the interface in 1E6E top pose (RMSD = 6.16 Å) is shown before (blue) and after (brown) 500 PELE

sampling steps; we can observe the formation of a new hydrogen bond along the PELE procedure. The overall conformational changes lower the interaction energy from E_{bind} (OPT) = -96.17 kcal/mol to E_{bind} (PELE) = -136.57 kcal/mol, as seen in Figure 49C.

The PELE method did not produce any improvement, however, in the 1BUH system. The techniques introduced here rely on a good initial sampling, which as shown above, is not the case for the 1BUH complex.

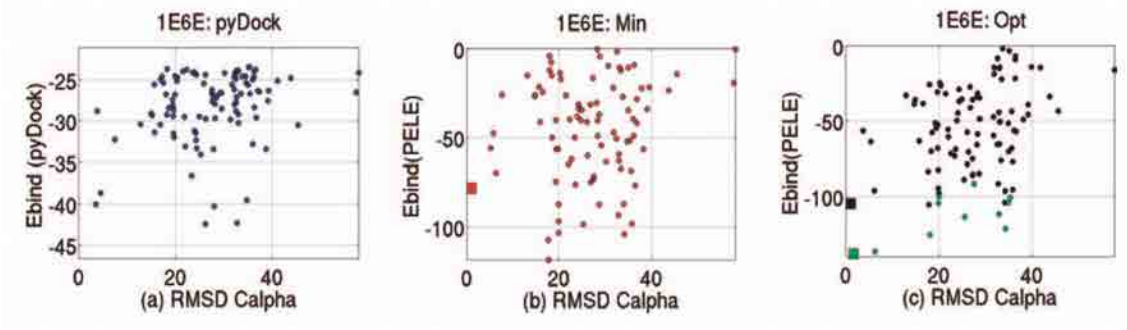


Figure 49. 1E6E scoring. Scoring of 100 poses for 1e6e protein–protein complexes. (a) 1E6E pyDock's scoring. (b) MIN procedure. (c) OPT(black) with PELE's refined top 10 (green) plus reference structure (square).

Applicability of the refinement techniques

As observed in previous (PELE) studies in protein–ligand induced fit interactions (Borrelli, Cossins et al. 2010), the techniques introduced here can only return a near native top score if the RMSD of the pose to the crystal reference is 5 Å (or less). Additionally, we want to emphasize that the all-atom refinement techniques are very sensitive to the structure. Missing residues such as loops, initial/terminal residues, and even side chains might introduce significant changes in the protein shape and its electrostatic content. The wrong placing of polar (carboxylic and amino) capping residues at the initial/terminal site will alter the placement of one full charge significantly. Similarly, cofactors (prosthetic groups) and ions have to be carefully analyzed. Furthermore, while less detailed scoring functions can tolerate wrong structures, for example with large protein–protein penetration or interface knot-like loops, the all atom force field score introduced here reacts drastically to such deficiencies. Interestingly, many of these wrong structures will introduce large electrostatic stabilizations as a result of a ''forced'' proximity of opposing charges introducing false positives. Such limitation (besides its larger computational cost) makes it difficult to apply these refinements to a large number of poses. In this initial application, we have limited the study to those complexes that had a near native within the first 100 poses, for which we could perform visual inspections. For this test set, we encountered very few of these clashes. When

increasing the set to the first 1000 poses per complex, however, we encountered dozens of bad structures introducing several false positives.

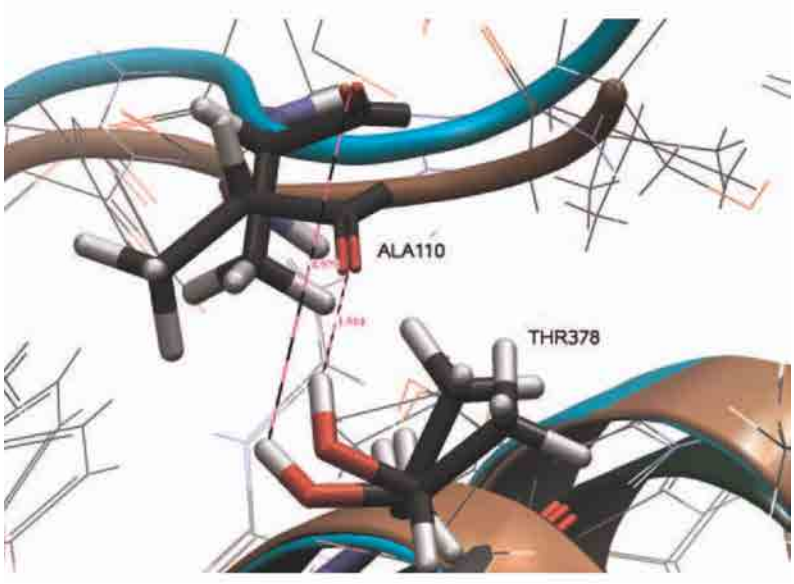


Figure 50. 1E6E conformational change. Detailed view of one of the conformational change introduced by PELE to 1E6E's top pose. Before PELE (blue ribbon), after PELE (brown ribbon).

We have expanded our study, however, to those complexes that had a near native conformation within the first 100 poses for other 12 complexes from the Benchmark 4.0. Thus, we increased the OPT study with an additional 12 complexes: 1r6q, 1oyv, 1n8o, 2ayo, 4cpa, 1z0k, 1xd3, 2vdb, 1rv6, 1jtg, 2oul, and 2b42. The results are again very satisfactory, improving the pyDock results considerably. In particular, as in the previous ten cases, the correlation between the scoring and the RMSD is significantly improved (Masone, Cabeza de Vaca et al. 2012) (see Figure 2 Appendix A).

5.2.2 Closure

We have presented a refinement protocol capable of rescoring near native poses in protein–protein interactions. An initial step involved an optimization of the hydrogen bond network, a procedure, with an approximate cost of 10 min per pose. Our results indicate a remarkable increase in the correlation between the interaction energy (score) and the RMSD to the reference crystal. The optimization cannot involve only the interface region, but needs to include the entire system. Optimization of the interface alone would maximize the interaction energy and not the total energy, introducing false positives. A second step, involving a deeper level of sampling, has been introduced with our in house code PELE. This technique involves considerable more computational resources and has been only applied, in a hierarchical scheme, to the 10 top poses obtained in the first step. The overall procedure indicates, as

observed in our previous studies in protein–ligand interactions (Borrelli, Cossins et al. 2010), that a standard classical force field with an implicit solvent is capable of successfully discriminating near native poses.

5.3 Applications of multiscale approach

5.3.1 Test case I: tryptogalinin

A salivary proteome-transcriptome project on the hard tick *Ixodes scapularis* revealed that Kunitz peptides are the most abundant salivary proteins. Ticks use Kunitz peptides (among other salivary proteins) to combat host defense mechanisms and to obtain a blood meal. Most of these Kunitz peptides, however, remain functionally uncharacterized, thus limiting our knowledge about their biochemical interactions.

In our study, we focused on the most abundant Kunitz group from the *I. scapularis sialome* project by Ribeiro et al. (Ribeiro, Alarcon-Chaidez et al. 2006): the monolaris group. We identified a Kunitz sequence that displays an unusual Cys motif when compared with the other monolaris and to previously reported Kunitz peptides. Since tick Kunitz peptides are known to inhibit serine proteases we performed an inhibitory screening demonstrating that this *I. scapularis* Kunitz inhibits several proteases as well as being a potent inhibitor of human skin β-tryptase (HSTb). We will, hereafter, refer to this *I. scapularis* Kunitz as tryptogalinin due to its high affinity for HSTb. Since the crystal structure of a similar protease inhibitor called TdPI and its complex with trypsin has been solved, we will use *in silico* methods (homology-based modelling, MD, and PELE) to elucidate the biophysical principles that determine tryptogalinin's protein fold, to predict its global tertiary structure and to hypothesize about its physicochemical interactions with serine proteases that account for its biochemical specificity – when compared with TdPI.

5.3.1.1 Calculations and discussions

Molecular Dynamics

MD simulations were performed with Desmond (Guo, Mohanty et al. 2010). The structures were solvated in an orthorhombic box, with a buffer solvent region of at least 10 Å. The system was neutralized, and an ionic force of 0.15 M was set. The default relaxation protocol in Desmond was used. The production run was in the NPT ensemble with a Nose-Hoover thermostat and a Martyna-Tobias-Klein barostat. The temperature was set to 300 K with a 2 fs time step, with a shake algorithm on hydrogen atoms and long range Ewald summation.

Protein-Protein Docking and Structural Refinement

We used several servers to identify a close to native complex between tryptogalinin and trypsin by performing a blind docking (i.e., we did not input any interacting residues between protease and inhibitor) using a trypsin monomer (PDB: 1TLD) (Berman, Westbrook et al. 2000). As a true positive control, the same procedure was performed using the monomers for TdPI (PDB: 2UUX), a modeled TdPI (using Modeller) and trypsin (PDB: 1TLD). We found that docking the monomers of TdPI (PDB: 2UUX) and trypsin (PDB: 1TLD) using the ClusPro 2.0 server (Comeau, Gatchell et al. 2004, Comeau, Gatchell et al. 2004, Kozakov, Brenke et al. 2006, Kozakov, Hall et al. 2010) generated a near to native crystal structure (6.3 Å RMSD) compared with other docking programs (>10 Å RMSD), such as PyDock (Gabb, Jackson et al. 1997, Cheng, Blundell et al. 2007) and FireDock (Duhovny, Nussinov et al. 2002, Schneidman-Duhovny, Inbar et al. 2005) – data not shown. Normally, poses of ~10 Å RMSD are considered to be a successful docking (Gabb, Jackson et al. 1997). Using the modeled TdPI and tryptogalinin, however, generated docked poses >15 Å than the native structure and introduced more false positives, even after the OPT refinement method (Masone, Cabeza de Vaca et al. 2012). We also attempted to indicate specific residues for the ClusPro server that come into contact upon binding (e.g., Lys-Asp), but this did not produce a proper docking pose, increased the number of false positives and reduced the number of generated poses – data not shown. All these shortcomings suggested a more robust technique must be applied in our docking methods.

In this sense, we have applied the multiscale approach to generate and identify tryptogalinin docking poses. Based on the TdPI-trypsin crystal (PDB: 2UUY), we added an 8 Å cutoff between Lys13 and Asp191 for tryptogalinin (restraint distance criterion). Following the global search step, we started from a configuration where both monomers are far apart, the algorithm first generates random large configurational jumps (up to 20 Å translation and 360º rotation) of the ligand (tryptogalinin) until the distance cutoff is satisfied. Then, the size of the random jumps decrease to perform 10.000 steps of local exploration (up to 3 Å and 5º). The overall procedure may be repeated several times. The distance cutoff, together with a steric clash screen, quickly populates the areas of interest (determined from the experimental information, etc.). Furthermore, new configurations are only accepted if five parameters related to relative positions between monomers differ by a range from any previous one. The parameters used to avoid the production of similar results are spherical coordinates of the center of mass of the ligand respect to the receptor and two spherical angles within the ligand. The overall procedure was capable of producing around 300.000 configurations in 10 hours on a single CPU.

All MC accepted steps within the cutoff constraint were clustered to 100 poses and converted back to all-atom models (keeping the initial atomic structure information). Following the

previous section, we refined the all-atom poses using OPT protocol that optimizes the entire hydrogen bond network by means of side chain sampling.

MD and intrinsic protein disorder reveal tryptogalinin's biochemical interactions

Tryptogalinin is an excellent candidate for refinement techniques using MD due to its small size and the presence of multiple Cys bridges; therefore, we refined the homologous tryptogalinin model with a 60 ns trajectory. As expected from a homology-modeled structure, we observed a rapid deviation from the initial conformation (~4 Å RMSD), followed by an equilibration. Figure 51A shows 100 equidistant structures for the last 40 ns and compares them to a TdPI simulation (Figure 51B; under the same conditions). We observe larger mobility in the L1 (as a result of the missing N-terminus disulfide bridge) and the L2 loop regions for tryptogalinin. Furthermore, this higher regional mobility results in the lysine 13 (Lys13) residue to explore a significantly larger area of space.

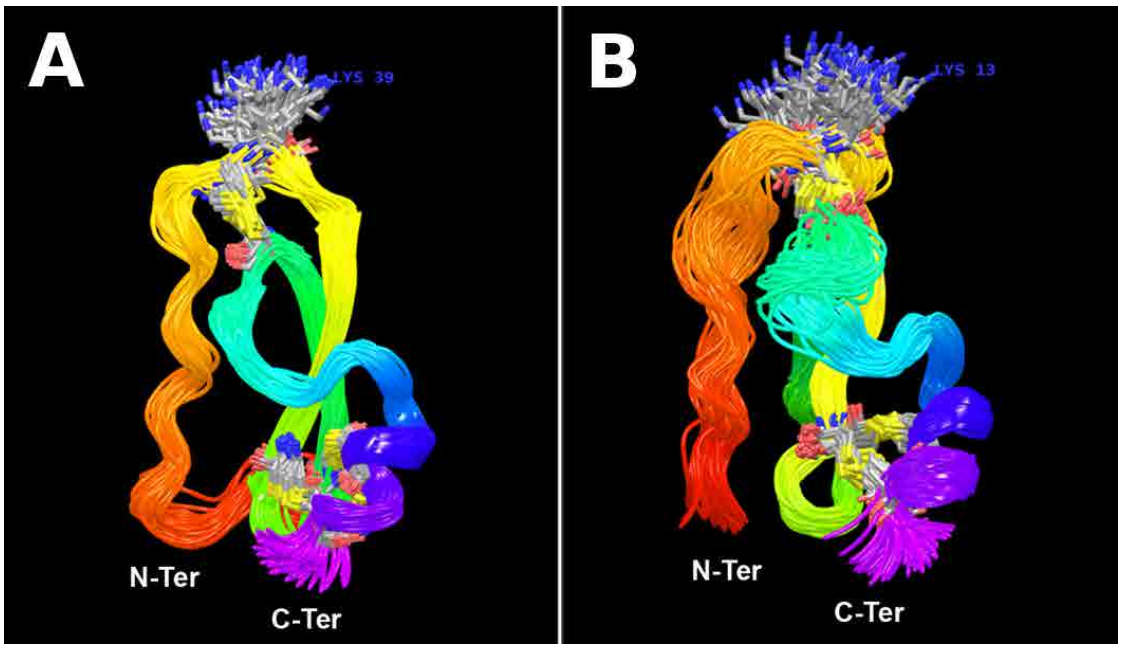


Figure 51. MD simulations. A total of 100 snapshots (i.e., conformations) during last 40 ns of MD for TdPI (Panel A) and tryptogalinin (Panel B).

Intrinsically disordered regions increase molecular recognition because of an ability to fold differently upon binding as well as possessing large interacting surfaces (Dunker, Brown et al. 2002). This may explain tryptogalinin's high affinity and multiple serine protease inhibition since part of its disorder extends from the N-terminus (L1) to the P1 interacting site (K13) compared with TdPI. Disorder is also predicted in the L2 region in proximity to the fourth Cys residue (the Cys forming the disulfide bond II with the P1 site). Such mobility,

however, might result into an induced fit recognition mechanism, therefore complicating any protein-protein docking simulations.

Tryptogalinin-trypsin docking verifies an induced fit recognition mechanism

Since the TdPI-trypsin crystallographic structure has been solved (PDB: 2UUY), we attempted to predict the tryptogalinin-trypsin complex by performing protein-protein docking. Initial blind docking of the homology model and of the last structure from the equilibration MD with ClusPro 2.0 (Comeau, Gatchell et al. 2004, Comeau, Gatchell et al. 2004, Kozakov, Brenke et al. 2006, Kozakov, Hall et al. 2010), PyDock (Gabb, Jackson et al. 1997, Cheng, Blundell et al. 2007) and FireDock (Duhovny, Nussinov et al. 2002, Schneidman-Duhovny, Inbar et al. 2005), did not produce any result <10 Å RMSD; the best scoring poses were located at RMSD distances >20 Å. Inspecting the generated poses it was apparent that Lys13 (and L1) was not able to approximate towards the trypsin active binding site. The distance between Lys13 of tryptogalinin and the Asp191 of the trypsin binding site was always >8 Å, whereas the distance between TdPI Lys39 and trypsin Asp191 is ~3 Å in the 2UUY crystallographic structure. Furthermore, when trying to superimpose the tryptogalinin model (or the MD equilibrated one) to TdPI in the 2UUY crystal, it was clear that the Lys conformation was significantly different from the one present in the TdPI crystal. Together with the MD results shown above, all these data point to a possible induced fit or a conformational selection mechanism for tryptogalinin.

To further test this hypothesis we proceeded by superimposing all the tryptogalinin MD snapshots to TdPI and found one structure with only 0.9 Å RMSD (for the Lys all-atom RMSD). Then we used this structure (Tryp2), plus the equilibrated MD model (Tryp1; with a ~4 Å Lys-Asp RMSD), for the following round of protein-protein docking studies with our multiscale approach. The top two panels in Figure 52A shows 300.000 MC steps for the CG exploration, where we biased tryptogalinin to the active site using the residue distance criteria explained below. Tryp2, the tryptogalinin MD conformation with better superimposition to TdPI, enters the active site reaching Lys13-Asp191 distances < 4 Å with a significant correlation between the CG binding energy and the RMSD. In agreement with the previous docking experiments, the equilibrated MD structure is not capable of entering the active site. From these CG results, we clustered 100 poses where we imposed the restraint distance criteria between Lys13 and Asp191 to be <8 Å, and refined them with all-atom models. The bottom two panels in Figure 52A show the all-atom binding energy. Clearly, Tryp2 produces again a significant correlation between the binding energy and the RMSD, using the 2UUY model as the reference, indicating that both proteins bind similarly.

Figure 52B shows a comparison between the TdPI crystallographic structure and the best tryptogalinin model after the all-atom refinement. The Lys orientation, α-helix and β-sheet placement significantly agree with that of the TdPI crystal complex. We should emphasize here that the only constraint in the simulation was the Lys-Asp distance, maintaining it below 8 Å (with a crystal value of 3.4 Å). Interestingly, the presence of a conformation in the tryptogalinin dynamics matching the TdPI bound structure, and producing similar bound complex with the best scoring, indicates a conformational selection binding mechanism.

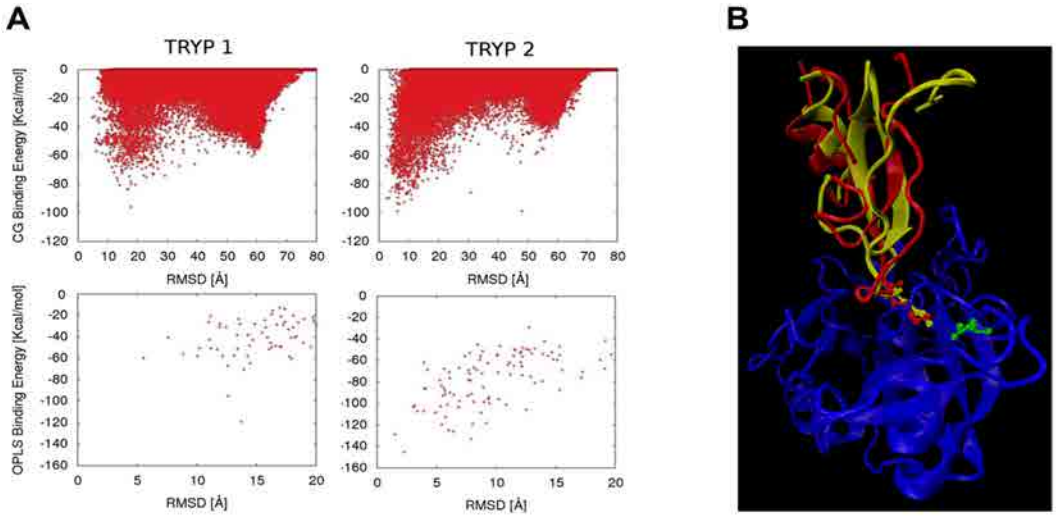


Figure 52. Coarse grain docking of refined tryptogalinin model and tryptogalinin-trypsin complex. Panel A, top two plots show the coarse grain binding against the RMSD for two tryptogalinin models, Tryp1: the last snapshot of a 62 ns equilibration, and Tryp2, the snapshot with best superimposition to TdPI (its complex with trypsin) (PDB: 2UUY). RMSD were obtained with respect to the superimposition of tryptogalinin to 2UUY. Bottom panels show the all-atom binding energy after clusterization of the coarse grain poses. Panel B, a comparison between the best all-atom model for tryptogalinin (yellow) and the complex TdPI-trypsin crystal structure (PDB: 2UUY) (red) depict significant binding similarities. Lys13 and 34, and Asp191 (green) are represented in ball and stick.

5.3.1.2 Closure

We hypothesize that the inhibitory profile of tryptogalinin is due to its intrinsic regional disorder, clearly shown in our MD simulations. Conventional docking methods proved to be inadequate due to the conformational selection binding mechanism of tryptogalinin. A theoretical combination of MD, superimposition to the TdPI crystal, CG Monte Carlo protein-protein docking, and all-atom refinement procedure, provided an adequate tryptogalinin-trypsin complex.

5.3.2 Test case II: a theoretical multiscale treatment of the FNR/Fd and FNR/Fld systems

In the photosynthetic electron transfer (ET) chain, two electrons transfer from Photosystem I to the flavin-dependent ferredoxin-NADP+ reductase (FNR) via two sequential independent

ferredoxin (Fd) electron carriers. In some algae and cyanobacteria (as Anabaena), under low iron conditions, flavodoxin (Fld) replaces Fd as single electron carrier. Extensive mutational studies have characterized the protein-protein interaction in FNR/Fd and FNR/Fld complexes. Interestingly, even though Fd and Fld share the interaction site on FNR, individual residues on FNR do not participate to the same extent in the interaction with each of the protein partners, pointing to different electron transfer mechanisms. Despite of extensive mutational studies, only FNR/Fd X-ray structures from Anabaena and Maize have been solved; structural data for FNR/Fld remains elusive An FNR/Fld bound model, however, has been proposed based on the high homology between two different domains of cytochrome P450 reductase (CPR) with FNR and Fld (Mayoral, Martínez-Júlvez et al. 2005). To check for the validity of this structure, and to model the FNR/Fld complex, we applied our multiscale modelling approach including CG and all-atom protein-protein docking. Moreover, we added the QM/MM e-pathway analysis and electronic coupling calculations, allowing for a molecular and electronic comprehensive analysis of the ET process in both complexes.

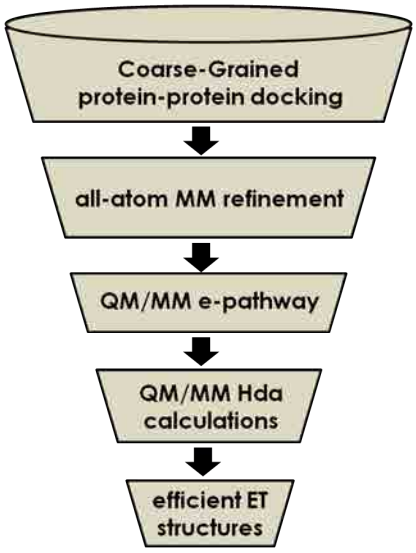


Figure 53. The 'funnel filtering' scheme to efficiently map the protein-protein ET mechanism.

5.3.2.1 Calculations and discussion

We protonated the system using pH 7 with PROPKA for Asp, Glu and His. Moreover, hydrogen bond networks were optimized producing better all atom conformations. We checked two implicit solvent models available in PELE based on a surface generalized born called SGBNP (Ghosh, Rapp et al. 1998, Gallicchio, Zhang et al. 2002) and OBC (Onufriev, Bashford et al. 2004) combined with an ionic strength model of Debye-Huckel (Edinger,

Cortis et al. 1997). Also, we tried weak and strong ionic strengths of 0.15 and 0.45 per each configuration to check how ionic strength affects the structures. Success rate parameter (on the know FNR/Fd complex) was the metric used to evaluate the results. It is expressed in percentage and is computed counting the number of structures in the top10 with an RMSD less than 10 Å, 5 Å and 2 Å.

FNR/Fd complex

In our initial attempt to apply our CG sampling for the FNR/Fd complex, where a reference crystal structure exists, we obtained good cofactor distances, in the range of 9.5-12.4 Å. Moreover, the reference X-ray structure was always lower in energy, -i.e. no false positives were produced. However, the CG sampling was not able to positively score near native structures with RMSD less than 4 Å see Figure 54A. Structural analysis of the low RMSD rejected structures showed negatively charged residues (GLU and ASP) at the binding interface in close proximity. These introduced a large repulsive interaction since our CG initial model kept them always deprotonated. PROPKA predictions, however clearly determine one of them to be protonated (in the complex), because of the electrostatic interaction between pair-wise negatively charged residues. To solve this problem, a discrete protonation criterion was implemented in the CG sampling to take into account possible pKa changes of surface negative residues upon complex formation. In particular, for each conformation (and before scoring takes place), if two negatively charged surface residues from protein A (FNR) and B (Fd or Fld) are within 6 Å, measured as side chain bead distance, the ligand residue gets protonated to its neutral state (in a similar approach to the one used by PROPKA). This procedure reduced repulsive interactions between negatively charged beads along the protein-protein interface eliminating false negatives. This modification of our CG sampling algorithm to take such effect into account resulted in a significant improvement of the CG sampling with acceptance of low energetic conformations with RMSD less than 4 Å, see Figure 54B. The near native conformation with a RMSD of 1.4 Å from the X-ray reference is now ranked as the best solution. Moreover, among the top 10-lowest energy solutions, we find four additional structures resembling the X-ray complex (with RMSD < 3 Å), together with two other distinct minima at 13 and 19 Å RMSD. Importantly, in the subsequent all-atom refinement, the funnel correlation between the binding energy and the RMSD against the native X-ray structure is better observed, see Figure 54C. Notice as well that the X-ray structure was also minimized, with a consequent small RMSD displacement of 1.0 Å. This result indicates a good correlation between CG and all-atom energy functions and validates the faster CG screening of the number of candidates to be scored by all-atom techniques.

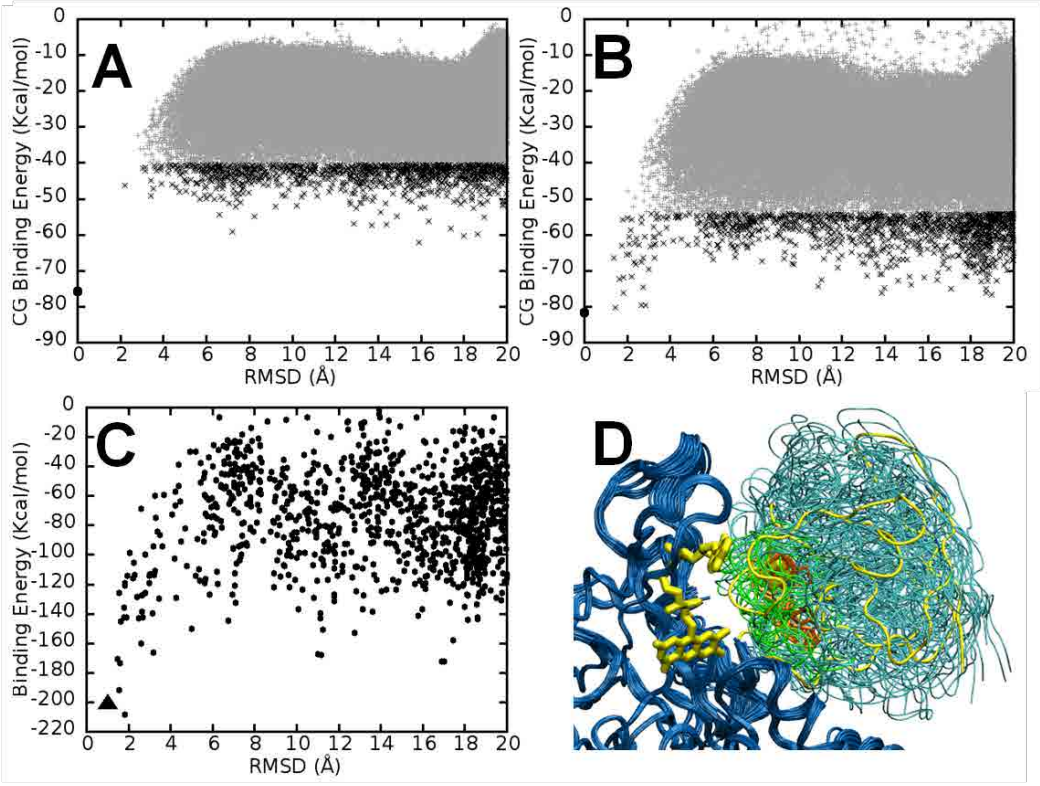


Figure 54. FNR/Fd complex sampling. Plots of CG docking binding energy versus RMSD to the reference X-Ray structure without (Panel A) and with (Panel B) re-protonation of surface charged residues. The reference is marked with big black dot at 0 Å RMSD. The top 1500 structures selected for the following all-atom refinements are underlined darker. Panel C, the plot of the top 1500 refined all-atom binding energy versus the same RMSD. The reference is marked with black triangle. Panel D, superimposition of 20 lowest energy structures representing each 1 Å RMSD window from the all-atom refinement. The reference structure is shown in yellow color. FNR protein is in dark blue, Fd protein is in cyan, FES of Fd is in orange, the loop-residue 40-49 of Fd is in green.

The lowest RMSD poses (< 4 Å) share the same interaction site as the X-ray one, forming strong hydrogen bonds between Fd:E94/E95 with FNR:K72 or FNR:K75, and Fd:D67/D69 with FNR:R16, as well as hydrophobic interactions between F65 on Fd with L76, L78 and V136 on FNR. These residues on FNR and Fd have been identified to be critical for protein-protein interactions by mutational studies (Martínez-Júlvez, Medina et al. 1999, Medina and Gómez-Moreno 2004, Medina 2009). Other conformational minima orient different negatively charged residues on Fd surface to interact with K72, K75 and R16 on FNR, such as Fd:D31/D36 or Fd:D62/D67/D69. Mutations at these Fd residues also produce a moderate effect on complex stability and ET with the reductase (Hurley, Morales et al. 2002). Notice that the 0-1 RMSD window structure corresponds to the minimized X-Ray one, not being a real prediction and only used for comparison. These candidate structures show multiple orientations of Fd binding on FNR, but all share Fd´s loop residues 40-49 at the interface with

FNR and thus have the redox distance between FAD and FES around 7 - 10 Å (measured between FAD:C8M•••FES:Fe1 atoms), see Figure 54D.

FNR/Fld complex

The CG sampling protocol fitted with the FNR/Fd complex was applied to model the interactions of the FNR/Fld complex, for which crystallographic structure is unknown. Contrary to the Fd complex, now we do not obtain a funnel-shaped correlation between binding energy and RMSD toward the reference homology model (Mayoral, Martínez-Júlvez et al. 2005), Figure 55A and 55B. The complexes obtained present conformations with FAD-FMN cofactor distances within a range of 4 – 10 Å (measured between the geometrical centers of FAD:C8M/C7M atoms and FMN:C8M/C7M atoms). The 10 lowest energy conformations are drastically different from the reference model, with 18 - 20 Å RMSD values, and present cofactor distances in the 5.5 - 8.5 Å range, not as short as in the reference model (4.3 Å). Upon all-atom refinement, the overall picture of predicted complexes does not change; the best pose is 19 Å RMSD from the homology reference structure, Figure 55C. Interestingly, the best energy complexes show more interface contacts involving experimentally identified critical charged residues on FNR: K72, K75 and R16 as well as key hydrophobic residues (L76, L78 and V136). Superposition of the lowest energy conformation at each 1Å RMSD intervals in the all-atom refinement is shown in Figure 55D. They present multiple binding orientations of Fld, having the FMN cofactor in a direct contact with FNR protein.

While the reference homology model has the shortest distance (by 1.5 Å) between FAD and FMN, it represents a significant higher energy pose. Interestingly, there are several alternative orientations, for instance, structures at RMSD of 9, 12, and 19 Å, which bring the FAD and FMN rings into close distances and are associated to lower complex energies, which may dominate ET. As in Fd, we selected the 20 lowest energy conformations for each 1 Å RMSD-window (1-20 Å, with the 0-1 structure corresponding to the minimized CPR-homology reference) for further QM/MM e-Pathway and electronic coupling calculations. In FNR/Fd, the QM/MM e-pathway results indicated a bridge-mediated ET mechanism through the Fd loop involving residues 40-49. Moreover, electronic coupling (Hda) calculations confirmed the active role of this loop in assisting the ET process. Importantly, Hda values were correlated with the RMSD to the X-ray complex and the redox centers distance. In the FNR/Fld complex, however, we had a rather different scenario. The strong correlation between Hda values and the cofactor distance, strongly suggested a direct ET mechanism

between FAD and FMN. Only conformations having significantly short distance (within a van de Waals contact between cofactors) resulted in large Hda values (Saen-oon, de Vaca et al. 2015) (see Appendix B).

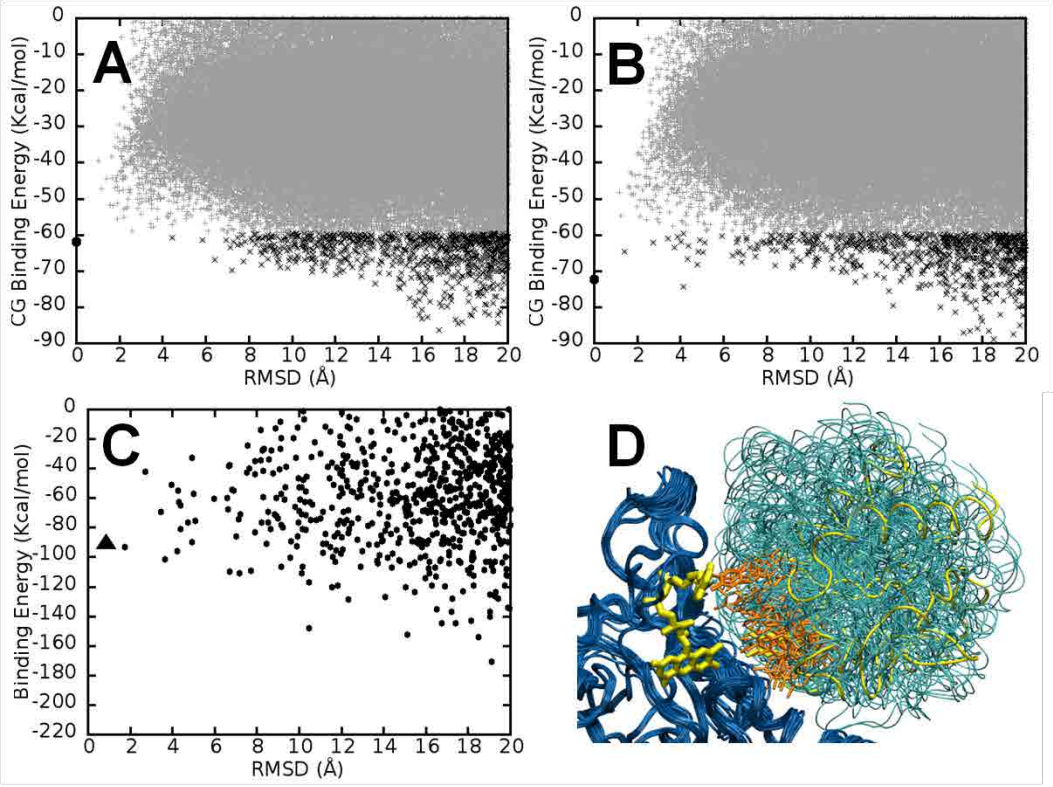


Figure 55. FNR/Fld complex sampling. Plots of CG docking binding energy versus RMSD to the reference homology structure without (Panel A) and with (Panel B) re-protonation of surface charged residues. The reference is marked with big black dot at 0 Å RMSD. The top 1500 structures selected for the following all-atom refinements are underlined darker. Panel C, the plot of all-atom refinement binding energy versus RMSD. The reference is marked with black triangle. Panel D, superimposition of 20 lowest energy structures representing each 1 Å RMSD window from the all atom refinement. The reference structure is shown in yellow. FNR protein is in dark blue, Fld protein is in cyan, FMN of Fld is in orange.

5.3.2.2 Closure

Comparing protein-protein interaction energies, both at the CG (Figure 54B–Figure 55B) and all-atom level (Figure 54C–Figure 55C), we see how the interaction in Fld is less specific than that in Fd. Fd presents a deeper minima that should dominate the protein-protein interaction. In addition, best interaction energies are predicted in the vicinity of the reference crystal; we remind here that no information on the crystal FNR/Fd structure (other than for comparison purpose) is used along the protein-protein sampling. Fld, on the contrary, has a larger range of orientations showing similar interaction energies. Such initial biophysical

analysis seems to agree with the less specific FNR/Fld interaction scenario proposed from mutational analyzes.

Chapter 6

Conclusions

- Inclusion of an specific force field and anisotropic network model for DNA simulations, combined with a newer implicit solvent, has shown an excellent agreement with standard molecular dynamics simulations generating a similar set of conformations for six representative DNA fragments.

- DNA-ligand exploration with the new PELE features has shown the capacity to explore the whole DNA surface and find the binding site for different DNA binders. The study of the distributions allowed us to identify the best DNA binders providing accurate estimations of the binding free energies for each ligand.

- The addition of a new implicit solvent has speed up the calculations and PELE still providing accurate results to study protein-ligand interactions and provide bound complexes in cases where crystal structure is not available.

- We have been able to simulate the intercalation of a ligand into different DNA fragments with PELE in a few hours.

- We have shown how the modified PELE algorithm to steer atoms can simulate the force-extension profile of different proteins, providing almost equal results than steered molecular dynamics in less computational time.

- We were able to generate the same rupture length distribution with the steering PELE implementations than the atomic force microscopy experiments performed by our collaborators for the azurin apo/holo protein.

- Our MCPRO modification to steer atoms has been applied to evaluate unfolding free energies of deca-alanine opening the door to future studies based on non-equilibrium simulations to evaluate free energies with Monte Carlo approaches.

- In this thesis, using the multiscale approach developed, we were able to generate and discriminate thousands of protein-protein conformations in an affordable time successfully.

LIST OF PUBLICATIONS

During this thesis nine papers have been produced: seven papers published, one under review and one manuscript almost finished. Here, there is the papers list where Israel Cabeza de Vaca Lopez have contributed. Articles with * have not been included in this thesis. Authors marked # contributed equally to the work.

Diego Masone, Israel Cabeza de Vaca, Carles Pons, Juan Fernandez Recio, Victor Guallar (2012). H-bond network optimization in protein-protein complexes: Are all-atom force field scores enough? *Proteins Structure Function and Bioinformatics: 80(3): 818-24*. (Impact factor: 2.63, 8 citations)

James J Valdés, Alexandra Schwarz, Israel Cabeza de Vaca, Eric Calvo, Joao H F Pedra, Victor Guallar, Michalis Kotsyfakis (2013). Tryptogalinin Is a Tick Kunitz Serine Protease Inhibitor with a Unique Intrinsic Disorder. *PLoS ONE: 8(5): e62562*. (Impact factor: 3.23, 13 citations)

Israel Cabeza de Vaca #, Maria F Lucas #, Ryoji Takahashi, Jaime Rubio-Martínez, Víctor Guallar (2014). Atomic Level Rendering of DNA-Drug Encounter. *Biophysical Journal: 106(2): 421-9*. (Impact factor 3.97, 5 citations)

Israel Cabeza de Vaca #,Giuseppina Andreotti #, Angelita Poziello, Maria Chiara Monti, Victor Guallar, Maria Vittoria Cubellis (2014). Conformational Response to Ligand Binding in Phosphomannomutase2. *Journal of Biological Chemistry; 289(50).* (Impact Factor 4.57, 2 citations)

Israel Cabeza de Vaca #, Marina I. Giannotti #, Juan Manuel Artés, Fausto Sanz, Victor Guallar, Pau Gorostiza (2015). Direct Measurement of the Nanomechanical Stability of a Redox Protein Active Site and Its Dependence upon Metal Binding. *The Journal of Physical Chemistry B; 119(36):12050-12058.* (Impact factor 3.30)

Suwipa Saen-Oon, Israel Cabeza de Vaca, Diego Masone, Milagros Medina, Victor Guallar (2015). A theoretical multiscale treatment of protein-protein electron transfer: The

ferredoxin/ferredoxin-NADP(+) reductase and flavodoxin/ferredoxin-NADP(+) reductase systems. *Biochimica et Biophysica Acta.* (Impact factor 4.66)

Israel Cabeza de Vaca #, Jana Kopečná #, Nathan B. P. Adams, Paul A. Davison, Amanda A. Brindley, C. Neil Hunter, Victor Guallar, Roman Sobotka (2015). Porphyrin Binding to Gun4 protein, Facilitated by a Flexible Loop, Controls Metabolite Flow through the Chlorophyll Biosynthetic Pathway. *Journal of Biological Chemistry*. (Impact factor 4.57)

Israel Cabeza de Vaca, Maria F. Lucas, Victor Guallar. A new Monte Carlo based technique to study DNA-ligand interactions. *Under minor revision at Journal of Chemical Theory and Computation*. (Impact factor 5.49)

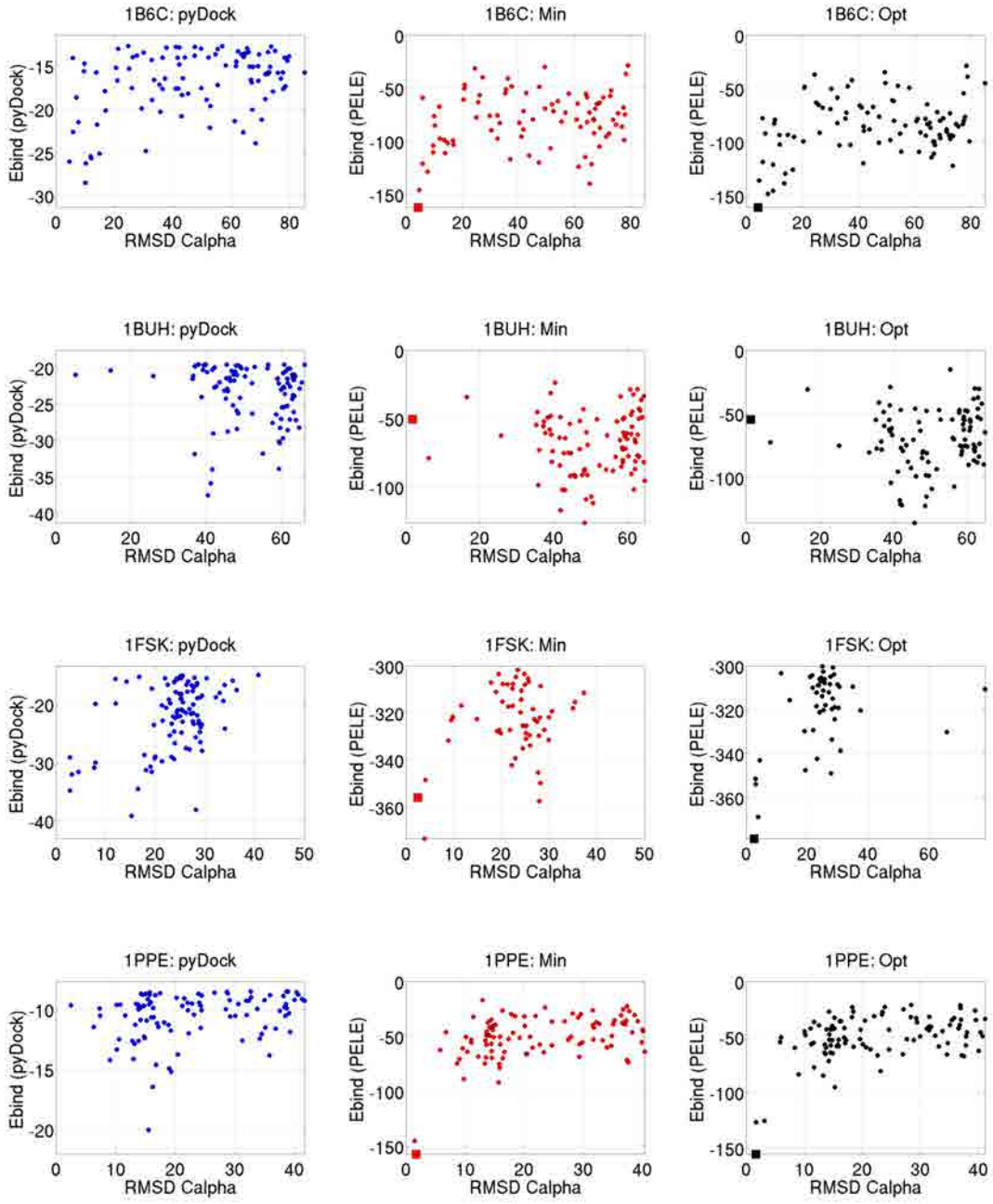
* Israel Cabeza de Vaca #, Sandra Acebes #, Victor Guallar. ecoupling server: A tool to compute and analyze electronic couplings. (*in preparation*)

INDEX

APPENDIX A

Multiscale scoring validation for protein–protein complexes.

Figure 1. Scoring of 100 poses for six selected protein–protein complexes from Benchmark 3.0. Left panel: pyDock, central panel: MIN refinement, and right panel: OPT refinement 0.45 ionic strength.



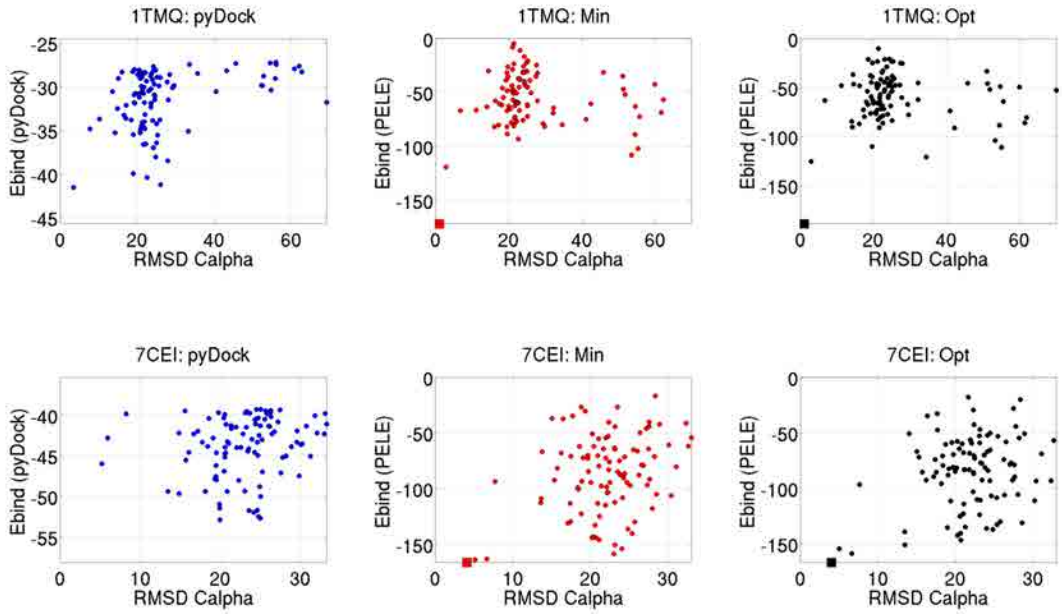
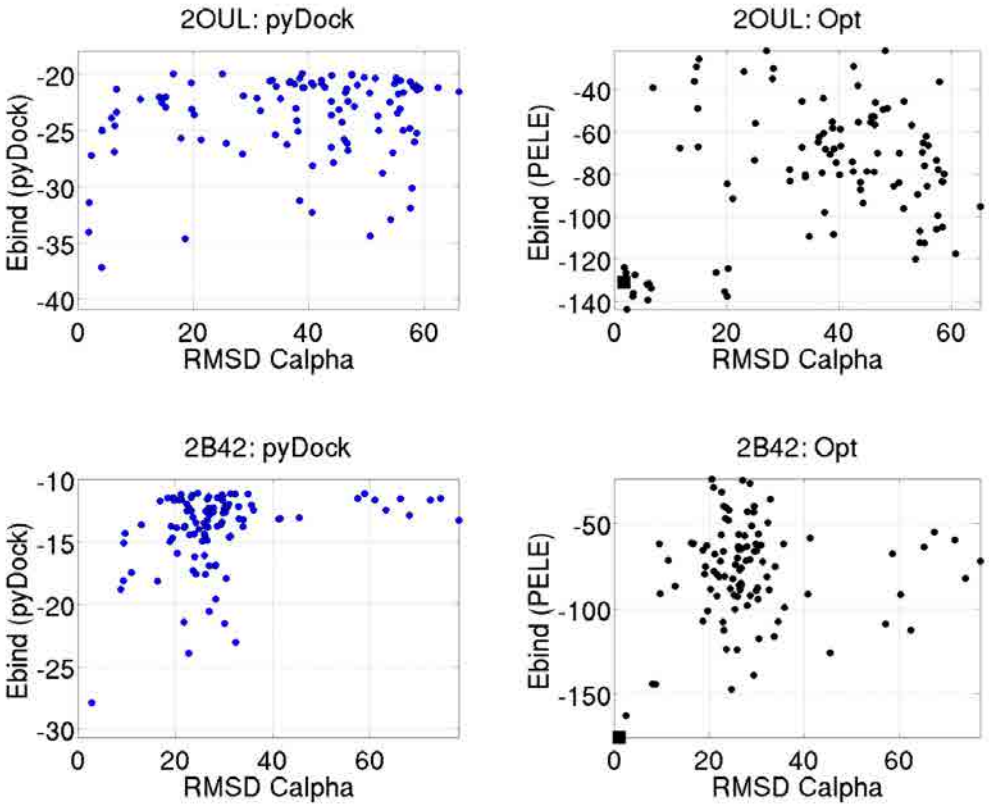
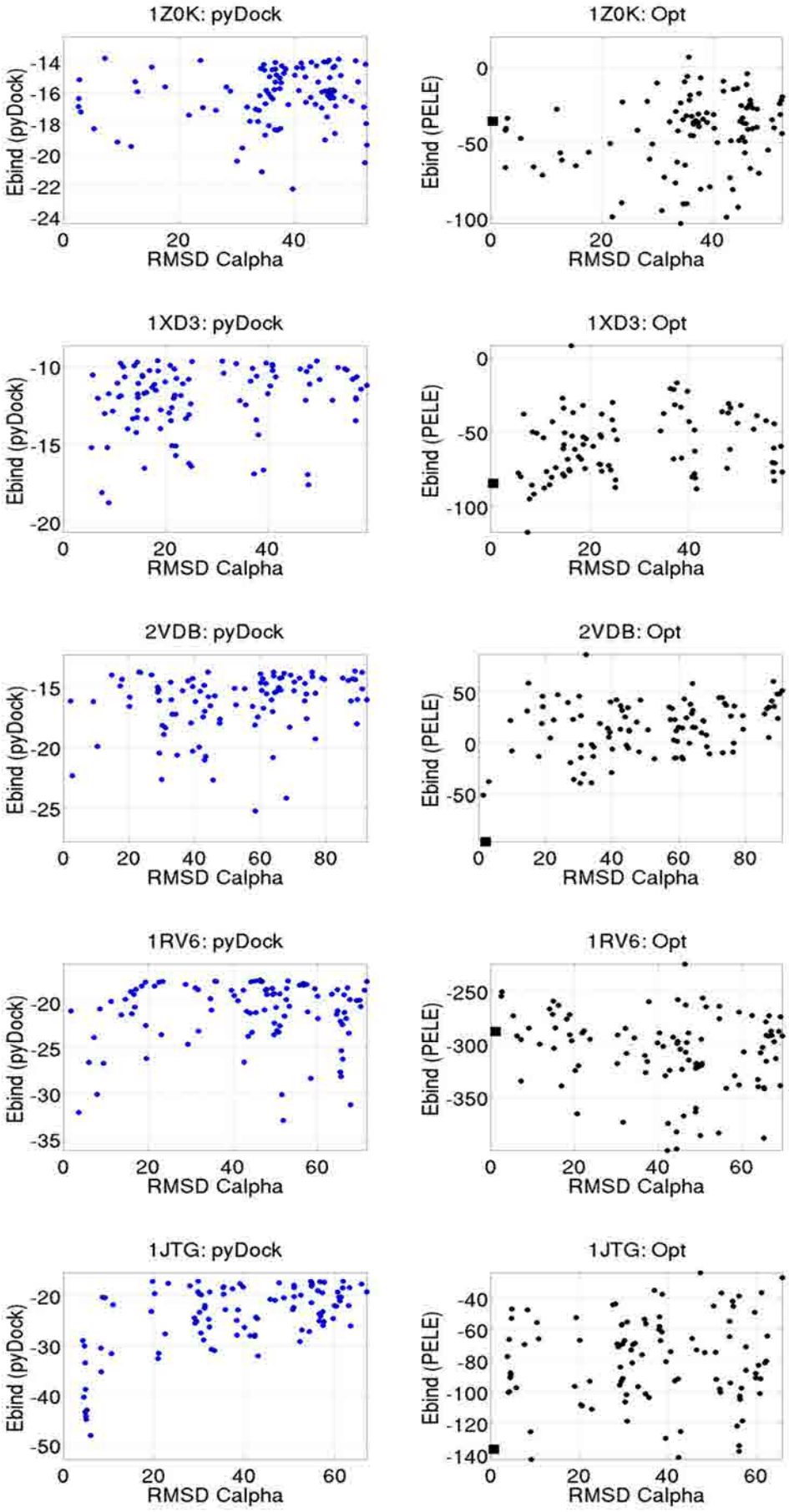
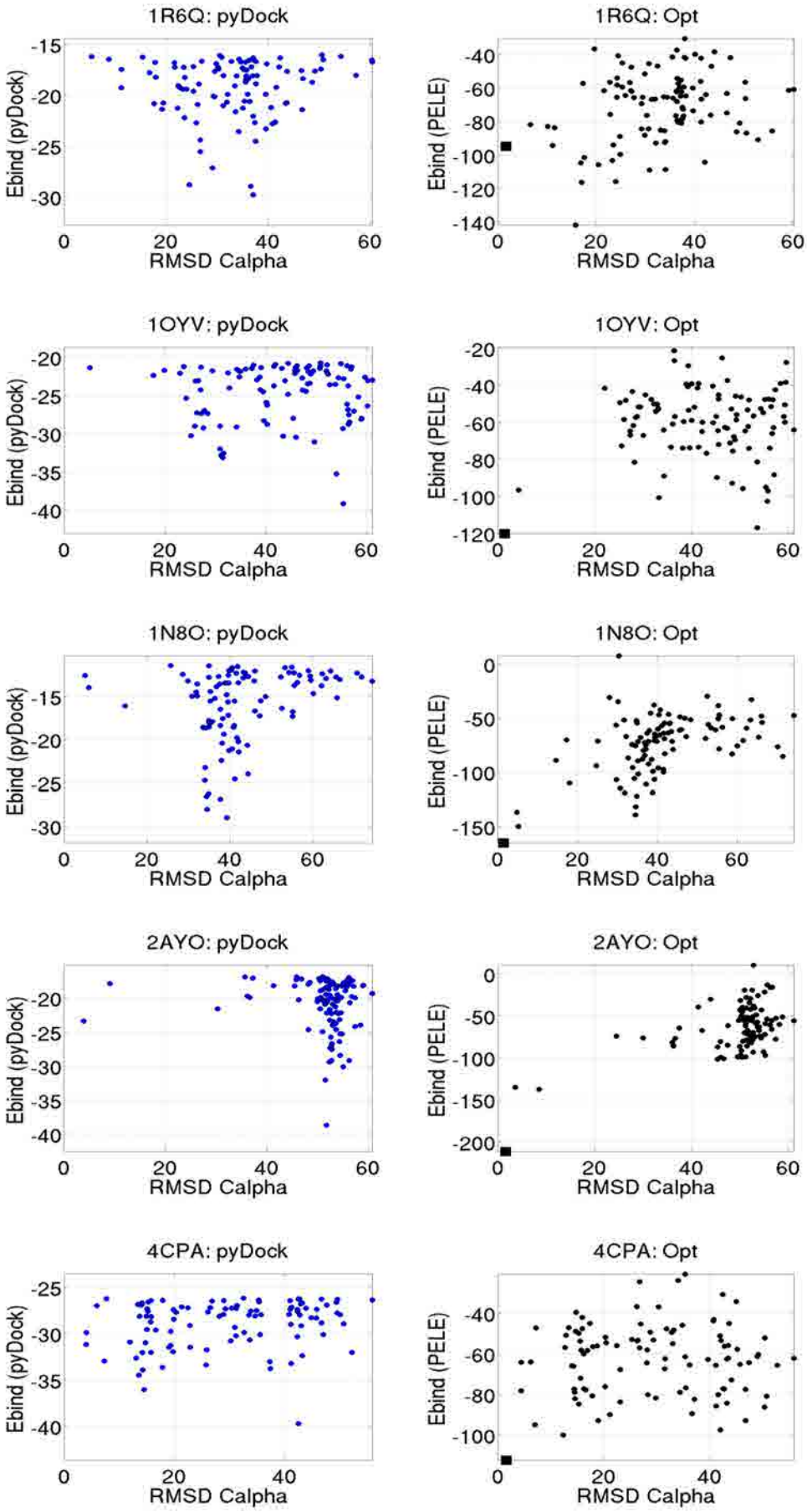


Figure 2. Scoring of 100 poses for twelve selected protein–protein complexes from Benchmark 4.0. Left panel: pyDock and right panel: OPT refinement with 0.45 ionic strength.



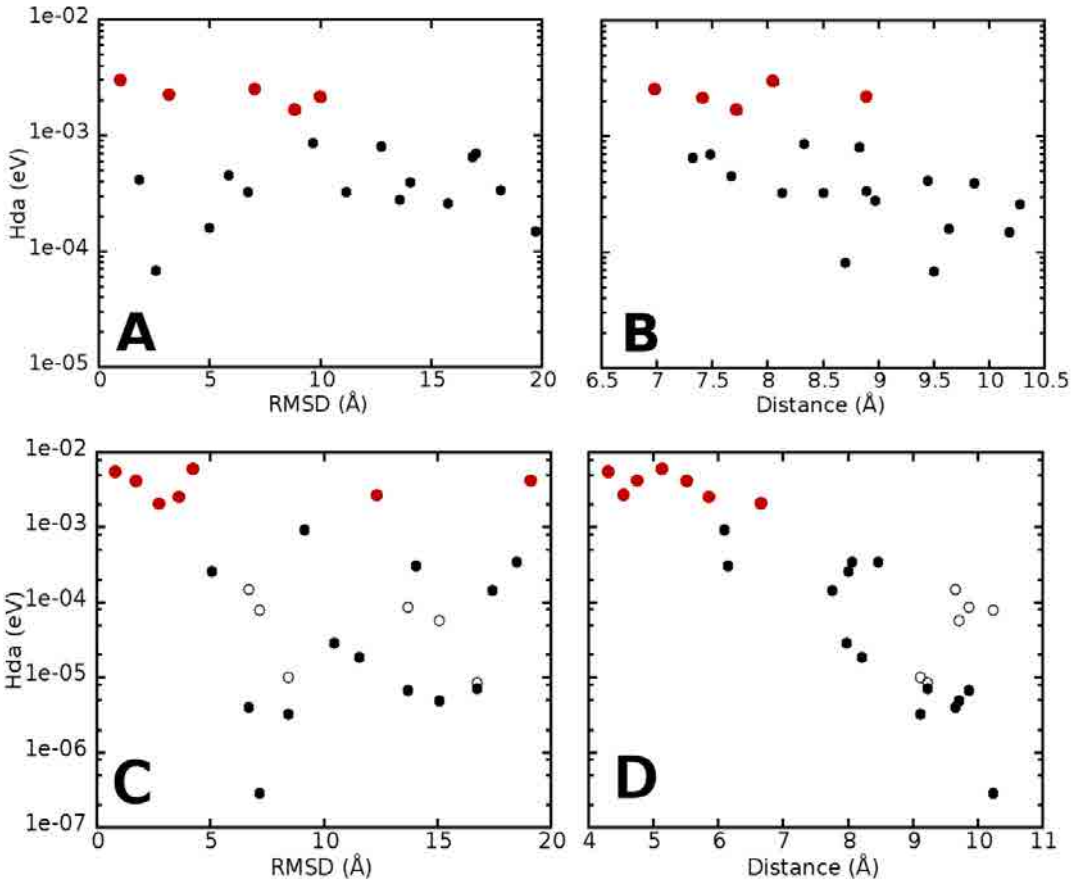

1Z0K: pyDock
Ebind (pyDock)
RMSD Calpha
1Z0K: Opt
Ebind (PELE)
RMSD Calpha
1XD3: pyDock
Ebind (pyDock)
RMSD Calpha
1XD3: Opt
Ebind (PELE)
RMSD Calpha
2VDB: pyDock
Ebind (pyDock)
RMSD Calpha
2VDB: Opt
Ebind (PELE)
RMSD Calpha
1RV6: pyDock
Ebind (pyDock)
RMSD Calpha
1RV6: Opt
Ebind (PELE)
RMSD Calpha
1JTG: pyDock
Ebind (pyDock)
RMSD Calpha
1JTG: Opt
Ebind (PELE)
RMSD Calpha


1R6Q: pyDock
Ebind (pyDock)
RMSD Calpha
1R6Q: Opt
Ebind (PELE)
RMSD Calpha
1OYV: pyDock
Ebind (pyDock)
RMSD Calpha
1OYV: Opt
Ebind (PELE)
RMSD Calpha
1N8O: pyDock
Ebind (pyDock)
RMSD Calpha
1N8O: Opt
Ebind (PELE)
RMSD Calpha
2AYO: pyDock
Ebind (pyDock)
RMSD Calpha
2AYO: Opt
Ebind (PELE)
RMSD Calpha
4CPA: pyDock
Ebind (pyDock)
RMSD Calpha
4CPA: Opt
Ebind (PELE)
RMSD Calpha

APPENDIX B

FNR/Fd and FNR/Fld electronic coupling values.

Logarithm plot of Hda values obtained from the 20 structures selected in the all-atom refinement versus the RMSD to the reference crystal and the redox distance. Panel A and B corresponds to the FNR/Fd system where redox distance is measured between FAD:C8M•••FES:Fe1. Panel C and D corresponds to the FNR/Fld system where redox distance is measured between the geometrical centers of FAD:C8M/C7M atoms and FMN:C8M/C7M atoms. Opened-circles indicate Hda values when including Y94 or W57 in the QM region. Relatively high Hda values are colored in red.



BIBLIOGRAPHY

(2015). Maestro. New York, Schrondiger, LLC.

Abriata, L. A., L. Banci, I. Bertini, S. Ciofi-Baffoni, P. Gkazonis, G. A. Spyroulias, A. J. Vila and S. Wang (2008). "Mechanism of CuA assembly." Nature chemical biology **4**(10): 599-601.

Abriata, L. A., A. J. Vila and M. Dal Peraro (2014). "Molecular dynamics simulations of apocupredoxins: insights into the formation and stabilization of copper sites under entatic control." JBIC Journal of Biological Inorganic Chemistry **19**(4-5): 565-575.

Adam, G. C., B. F. Cravatt and E. J. Sorensen (2001). "Profiling the specific reactivity of the proteome with non-directed activity-based probes." Chemistry & biology **8**(1): 81-95.

Adhikari, N. D., R. Orler, J. Chory, J. E. Froehlich and R. M. Larkin (2009). "Porphyrins promote the association of GENOMES UNCOUPLED 4 and a Mg-chelatase subunit with chloroplast membranes." Journal of Biological Chemistry **284**(37): 24783-24796.

Alderden, R. A., M. D. Hall and T. W. Hambley (2006). "The discovery and development of cisplatin." Journal of chemical education **83**(5): 728.

Andreotti, G., I. C. de Vaca, A. Poziello, M. C. Monti, V. Guallar and M. V. Cubellis (2014). "Conformational Response to Ligand Binding in Phosphomannomutase2 INSIGHTS INTO INBORN GLYCOSYLATION DISORDER." Journal of Biological Chemistry **289**(50): 34900-34910.

Arfken, G. (1985). "The method of steepest descents." Mathematical methods for physicists **3**: 428-436.

Arkhipov, A., P. L. Freddolino, K. Imada, K. Namba and K. Schulten (2006). "Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum." Biophysical journal **91**(12): 4589-4597.

Arkhipov, A., P. L. Freddolino and K. Schulten (2006). "Stability and dynamics of virus capsids described by coarse-grained modeling." Structure **14**(12): 1767-1777.

Bakan, A., L. M. Meireles and I. Bahar (2011). "ProDy: protein dynamics inferred from theory and experiments." Bioinformatics **27**(11): 1575-1577.

Baker, N. A. (2005). "Improving implicit solvent simulations: a Poisson-centric view." Current opinion in structural biology **15**(2): 137-143.

Baker, N. A., D. Sept, M. J. Holst and J. A. McCammon (2001). "The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers." IBM Journal of Research and Development **45**(3.4): 427-438.

Ballester, P. J. and J. B. Mitchell (2010). "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking." Bioinformatics **26**(9): 1169-1175.

Banci, L., I. Bertini, F. Cantini, I. C. Felli, L. Gonnelli, N. Hadjiliadis, R. Pierattelli, A. Rosato and P. Voulgaris (2006). "The Atx1-Ccc2 complex is a metal-mediated protein-protein interaction." Nature chemical biology **2**(7): 367-368.

Barth, P., T. Alber and P. Harbury (2007). "Accurate, conformation-dependent predictions of solvent effects on protein ionization constants." Proceedings of the National Academy of Sciences **104**(12): 4898-4903.

Basdevant, N., D. Borgis and T. Ha-Duong (2007). "A coarse-grained protein-protein potential derived from an all-atom force field." The Journal of Physical Chemistry B **111**(31): 9390-9399.

Basdevant, N., D. Borgis and T. Ha-Duong (2012). "Modeling protein–protein recognition in solution using the coarse-grained force field SCORPION." Journal of Chemical Theory and Computation **9**(1): 803-813.

Berendsen, H. (1987). "J. C.; Grigera, JR; Straatsma." TP Journal of Physical Chemistry.

Berendsen, H. J., J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. Haak (1984). "Molecular dynamics with coupling to an external bath." The Journal of chemical physics **81**(8): 3684-3690.
Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The protein data bank." Nucleic acids research **28**(1): 235-242.
Bhattacharyya, D., S. Ramachandran, S. Sharma, W. Pathmasiri, C. L. King, I. Baskerville-Abraham, G. Boysen, J. A. Swenberg, S. L. Campbell, N. V. Dokholyan and S. G. Chaney (2011). "Flanking bases influence the nature of DNA distortion by platinum 1,2-intrastrand (GG) cross-links." PLoS One **6**(8): e23582.
Binder, K. and D. W. Heermann (2010). Monte Carlo Methods for the Sampling of Free Energy Landscapes. Monte Carlo Simulation in Statistical Physics, Springer**:** 153-174.
Binnig, G., C. F. Quate and C. Gerber (1986). "Atomic force microscope." Physical review letters **56**(9): 930.
Boas, F. E. and P. B. Harbury (2008). "Design of protein–ligand binding based on the molecular-mechanics energy model." Journal of molecular biology **380**(2): 415-424.
Boer, D. R., A. Canals and M. Coll (2009). "DNA-binding drugs caught in action: the latest 3D pictures of drug-DNA complexes." Dalton Transactions(3): 399-414.
Bonomi, M., D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci and R. A. Broglia (2009). "PLUMED: A portable plugin for free-energy calculations with molecular dynamics." Computer Physics Communications **180**(10): 1961-1972.
Borrelli, K. W., B. Cossins and V. Guallar (2010). "Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility." Journal of computational chemistry **31**(6): 1224-1235.
Borrelli, K. W., A. Vitalis, R. Alcantara and V. Guallar (2005). "PELE: protein energy landscape exploration. A novel Monte Carlo based technique." Journal of Chemical Theory and Computation **1**(6): 1304-1311.
Boulikas, T. and M. Vougiouka (2004). "Recent clinical trials using cisplatin, carboplatin and their combination chemotherapy drugs (review)." Oncology reports **11**(3): 559-595.
Bowers, K. J., E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. Gregersen, J. L. Klepeis, I. Kolossvary, M. Moraes and F. D. Sacerdoti (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. SC 2006 Conference, Proceedings of the ACM/IEEE, IEEE.
Brocchieri, L. and S. Karlin (2005). "Protein length in eukaryotic and prokaryotic proteomes." Nucleic acids research **33**(10): 3390-3400.
Buch, I., T. Giorgino and G. De Fabritiis (2011). "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations." Proceedings of the National Academy of Sciences **108**(25): 10184-10189.
Bullard, B., C. Ferguson, A. Minajeva, M. C. Leake, M. Gautel, D. Labeit, L. Ding, S. Labeit, J. Horwitz and K. R. Leonard (2004). "Association of the chaperone αB-crystallin with titin in heart muscle." Journal of Biological Chemistry **279**(9): 7917-7924.
Calimet, N., M. Schaefer and T. Simonson (2001). "Protein molecular dynamics with the generalized Born/ACE solvent model." Proteins: Structure, Function, and Bioinformatics **45**(2): 144-158.
Carrion-Vazquez, M., H. Li, H. Lu, P. E. Marszalek, A. F. Oberhauser and J. M. Fernandez (2003). "The mechanical stability of ubiquitin is linkage dependent." Nature Structural & Molecular Biology **10**(9): 738-743.
Case, D., T. Darden, T. E. Cheatham III, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang and K. Merz (2012). "AMBER 12." University of California, San Francisco **1**(3).
Case, D. and P. Kollman (2012). "AmberTools 12." University of California, San Francisco.
Case, D. A., T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods (2005). "The Amber biomolecular simulation programs." Journal of computational chemistry **26**(16): 1668-1688.

Case, D. A., T. Darden, T. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. Walker, W. Zhang and K. Merz (2010). Amber 11, University of California.
Chaires, J. B., J. E. Herrera and M. J. Waring (1990). "Preferential binding of daunomycin to 5'TACG and 5'TAGC sequences revealed by footprinting titration experiments." Biochemistry **29**(26): 6145-6153.
Cheatham III, T. E., P. Cieplak and P. A. Kollman (1999). "A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat." Journal of Biomolecular Structure and Dynamics **16**(4): 845-862.
Cheatham III, T. E., J. Srinivasan, D. A. Case and P. A. Kollman (1998). "Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution." Journal of Biomolecular Structure and Dynamics **16**(2): 265-280.
Chen, K.-X., N. Gresh and B. Pullman (1985). "A theoretical investigation on the sequence selective binding of daunomycin to double-stranded polynucleotides." Journal of Biomolecular Structure and Dynamics **3**(3): 445-466.
Chen, X., H. Pu, X. Wang, W. Long, R. Lin, L. Liu, Q. Chen, D. Fan, G. Wang and X. Wang (2015). "Crystal structures of GUN4 in complex with porphyrins." Molecular plant.
Cheng, T. M. K., T. L. Blundell and J. Fernandez-Recio (2007). "pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking." Proteins: Structure, Function, and Bioinformatics **68**(2): 503-515.
Comeau, S. R., D. W. Gatchell, S. Vajda and C. J. Camacho (2004). "ClusPro: a fully automated algorithm for protein–protein docking." Nucleic acids research **32**(suppl 2): W96-W99.
Comeau, S. R., D. W. Gatchell, S. Vajda and C. J. Camacho (2004). "ClusPro: an automated docking and discrimination method for the prediction of protein complexes." Bioinformatics **20**(1): 45-50.
Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman (1995). "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." Journal of the American Chemical Society **117**(19): 5179-5197.
Cossins, B. P., A. Hosseini and V. Guallar (2012). "Exploration of protein conformational change with PELE and meta-dynamics." Journal of Chemical Theory and Computation **8**(3): 959-965.
Darden, T., D. York and L. Pedersen (1993). "Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems." The Journal of chemical physics **98**(12): 10089-10092.
Davison, P. A., H. L. Schubert, J. D. Reid, C. D. Iorg, A. Heroux, C. P. Hill and C. N. Hunter (2005). "Structural and biochemical characterization of Gun4 suggests a mechanism for its role in chlorophyll biosynthesis." Biochemistry **44**(21): 7603-7612.
DIMARCO, A., M. Gaetani, L. Dorigotti, M. Soldati and O. Bellini (1964). "DAUNOMYCIN: A NEW ANTIBIOTIC WITH ANTITUMOR ACTIVITY." Cancer chemotherapy reports. Part 1 **38**: 31-38.
Dominguez, C., R. Boelens and A. M. Bonvin (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." Journal of the American Chemical Society **125**(7): 1731-1737.
Dominy, B. N. and C. L. Brooks (1999). "Development of a generalized Born model parametrization for proteins and nucleic acids." The Journal of Physical Chemistry B **103**(18): 3765-3773.
Doruker, P., A. R. Atilgan and I. Bahar (2000). "Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor." Proteins: Structure, Function, and Bioinformatics **40**(3): 512-524.
Duan, Y., C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo and T. Lee (2003). "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations." Journal of computational chemistry **24**(16): 1999-2012.
Dudek, M. J. and J. W. Ponder (1995). "Accurate modeling of the intramolecular electrostatic energy of proteins." Journal of computational chemistry **16**(7): 791-816.

Duhovny, D., R. Nussinov and H. J. Wolfson (2002). Efficient unbound docking of rigid molecules. Algorithms in bioinformatics, Springer**:** 185-200.
Dunker, A. K., C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic (2002). "Intrinsic disorder and protein function." Biochemistry **41**(21): 6573-6582.
Earl, D. J. and M. W. Deem (2005). "Parallel tempering: Theory, applications, and new perspectives." Physical Chemistry Chemical Physics **7**(23): 3910-3916.
Edinger, S. R., C. Cortis, P. S. Shenkin and R. A. Friesner (1997). "Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation." The Journal of Physical Chemistry B **101**(7): 1190-1197.
Freddolino, P. L., F. Liu, M. Gruebele and K. Schulten (2008). "Ten-microsecond molecular dynamics simulation of a fast-folding WW domain." Biophysical journal **94**(10): L75-L77.
Friesner, R. A., J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley and J. K. Perry (2004). "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." Journal of medicinal chemistry **47**(7): 1739-1749.
Friesner, R. A., R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz (2006). "Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes." Journal of medicinal chemistry **49**(21): 6177-6196.
Fu, Y., H.-C. T. Tsui, K. E. Bruce, L.-T. Sham, K. A. Higgins, J. P. Lisher, K. M. Kazmierczak, M. J. Maroney, C. E. Dann III and M. E. Winkler (2013). "A new structural paradigm in copper resistance in Streptococcus pneumoniae." Nature chemical biology **9**(3): 177-183.
Gabb, H. A., R. M. Jackson and M. J. Sternberg (1997). "Modelling protein docking using shape complementarity, electrostatics and biochemical information." Journal of molecular biology **272**(1): 106-120.
Gallicchio, E. and R. M. Levy (2004). "AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling." Journal of computational chemistry **25**(4): 479-499.
Gallicchio, E., L. Y. Zhang and R. M. Levy (2002). "The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators." Journal of computational chemistry **23**(5): 517-529.
Ghosh, A., C. S. Rapp and R. A. Friesner (1998). "Generalized Born model based on a surface integral formulation." The Journal of Physical Chemistry B **102**(52): 10983-10990.
Ghosh, A. K., Z. L. Dawson and H. Mitsuya (2007). "Darunavir, a conceptually new HIV-1 protease inhibitor for the treatment of drug-resistant HIV." Bioorganic & medicinal chemistry **15**(24): 7576-7580.
Giannotti, M. I., I. Cabeza de Vaca, J. M. Artés, F. Sanz, V. Guallar and P. Gorostiza (2015). "Direct Measurement of the Nanomechanical Stability of a Redox Protein Active Site and Its Dependence upon Metal Binding." Journal of Physical Chemistry B.
Gilad, Y. and H. Senderowitz (2013). "Docking studies on DNA intercalators." Journal of Chemical Information and Modeling **54**(1): 96-107.
Girotto, S., L. Cendron, M. Bisaglia, I. Tessari, S. Mammi, G. Zanotti and L. Bubacco (2014). "DJ-1 is a copper chaperone acting on SOD1 activation." Journal of Biological Chemistry **289**(15): 10887-10899.
Goldacre, R., A. Loveless and W. Ross (1949). "Mode of production of chromosome abnormalities by the nitrogen mustards: the possible role of cross-linking." Nature **163**(4148): 667-669.
Goldstein, G., M. Scheid, U. Hammerling, D. Schlesinger, H. Niall and E. Boyse (1975). "Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells." Proceedings of the National Academy of Sciences **72**(1): 11-15.
Götz, A. W., M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker (2012). "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born." Journal of chemical theory and computation **8**(5): 1542-1555.

Graham, C., T. Kurtz, S. Meleard, P. Protter, M. Pulvirenti and D. Talay (1996). "Probabilistic numerical methods for partial differential equations: elements of analysis. 1996." Lecture Notes in Mathematics.
Grosdidier, A., V. Zoete and O. Michielin (2011). "SwissDock, a protein-small molecule docking web service based on EADock DSS." Nucleic acids research **39**(suppl 2): W270-W277.
Guallar, V., M. Jacobson, A. McDermott and R. A. Friesner (2004). "Computational modeling of the catalytic reaction in triosephosphate isomerase." Journal of molecular biology **337**(1): 227-239.
Guo, Z., U. Mohanty, J. Noehre, T. K. Sawyer, W. Sherman and G. Krilov (2010). "Probing the α-Helical Structural Stability of Stapled p53 Peptides: Molecular Dynamics Simulations and Analysis." Chemical biology & drug design **75**(4): 348-359.
Hammersley, J. M. (1960). "Monte Carlo methods for solving multivariable problems." Annals of the New York Academy of Sciences **86**(3): 844-874.
Hannon, M. J. (2007). "Supramolecular DNA recognition." Chemical Society Reviews **36**(2): 280-295.
Hawkins, G. D., C. J. Cramer and D. G. Truhlar (1995). "Pairwise solute descreening of solute charges from a dielectric medium." Chemical Physics Letters **246**(1): 122-129.
Hawkins, G. D., C. J. Cramer and D. G. Truhlar (1996). "Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium." The Journal of Physical Chemistry **100**(51): 19824-19839.
Hernández-Ortega, A., F. Lucas, P. Ferreira, M. Medina, V. Guallar and A. T. Martínez (2011). "Modulating O2 Reactivity in a Fungal Flavoenzyme INVOLVEMENT OF ARYL-ALCOHOL OXIDASE PHE-501 CONTIGUOUS TO CATALYTIC HISTIDINE." Journal of Biological Chemistry **286**(47): 41105-41114.
Hernández-Ortega, A., F. t. Lucas, P. Ferreira, M. Medina, V. Guallar and A. T. Martínez (2012). "Role of active site histidines in the two half-reactions of the aryl-alcohol oxidase catalytic cycle." Biochemistry **51**(33): 6595-6608.
Hernández-Ortega, A., P. Ferreira, P. Merino, M. Medina, V. Guallar and A. T. Martínez (2012). "Stereoselective Hydride Transfer by Aryl-Alcohol Oxidase, a Member of the GMC Superfamily." ChemBioChem **13**(3): 427-435.
Hershko, A., E. Eytan, A. Ciechanover and A. Haas (1982). "Immunochemical analysis of the turnover of ubiquitin-protein conjugates in intact cells. Relationship to the breakdown of abnormal proteins." Journal of Biological Chemistry **257**(23): 13964-13970.
Hess, B., C. Kutzner, D. Van Der Spoel and E. Lindahl (2008). "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation." Journal of chemical theory and computation **4**(3): 435-447.
Honda, S., K. Yamasaki, Y. Sawada and H. Morii (2004). "10 residue folded peptide designed by segment statistics." Structure **12**(8): 1507-1518.
Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling (2006). "Comparison of multiple Amber force fields and development of improved protein backbone parameters." Proteins: Structure, Function, and Bioinformatics **65**(3): 712-725.
Hosseini, A., M. Brouk, M. F. Lucas, F. Glaser, A. Fishman and V. Guallar (2014). "Atomic picture of ligand migration in toluene 4-monooxygenase." The Journal of Physical Chemistry B **119**(3): 671-678.
Huo, S., I. Massova and P. A. Kollman (2002). "Computational alanine scanning of the 1: 1 human growth hormone–receptor complex." Journal of computational chemistry **23**(1): 15-27.
Hurley, J. K., R. Morales, M. Martınez-Júlvez, T. B. Brodie, M. Medina, C. Gómez-Moreno and G. Tollin (2002). "Structure–function relationships in Anabaena ferredoxin/ferredoxin: NADP+ reductase electron transfer: insights from site-directed mutagenesis, transient absorption spectroscopy and X-ray crystallography." Biochimica et Biophysica Acta (BBA)-Bioenergetics **1554**(1): 5-21.
Hurley, L. H. (2002). "DNA and its associated processes as targets for cancer therapy." Nature Reviews Cancer **2**(3): 188-200.

Hwang, H., T. Vreven, J. Janin and Z. Weng (2010). "Protein–protein docking benchmark version 4.0." Proteins: Structure, Function, and Bioinformatics **78**(15): 3111-3114.
Imanishi, J., Y. Morita, E. Yoshimi, K. Kuroda, T. Masunaga, K. Yamagami, M. Kuno, E. Hamachi, S. Aoki and F. Takahashi (2011). "Pharmacological profile of FK881 (ASP6537), a novel potent and selective cyclooxygenase-1 inhibitor." Biochemical pharmacology **82**(7): 746-754.
Israelewitz, B., M. Gao and K. Schulten (2001). "Steered molecular dynamics and mechanical functions of proteins." Current opinion in structural biology **11**(2): 224-230.
Ivani, I., Dans, P.D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrió, P., Goñi, R., Balaceanu, A. (2015). "ParmBSC1: A Refined Force-Field for DNA simulations. ." Nature methods **In press**.
Jacobson, M. P., R. A. Friesner, Z. Xiang and B. Honig (2002). "On the role of the crystal environment in determining protein side-chain conformations." Journal of molecular biology **320**(3): 597-608.
Jacobson, M. P., D. L. Pincus, C. S. Rapp, T. J. Day, B. Honig, D. E. Shaw and R. A. Friesner (2004). "A hierarchical approach to all-atom protein loop prediction." Proteins: Structure, Function, and Bioinformatics **55**(2): 351-367.
Jarzynski, C. (1997). "Nonequilibrium equality for free energy differences." Physical Review Letters **78**(14): 2690.
Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "Comparison of simple potential functions for simulating liquid water." The Journal of chemical physics **79**(2): 926-935.
Jorgensen, W. L., D. S. Maxwell and J. Tirado-Rives (1996). "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids." Journal of the American Chemical Society **118**(45): 11225-11236.
Jorgensen, W. L. and J. Tirado-Rives (1988). "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin." Journal of the American Chemical Society **110**(6): 1657-1666.
Jorgensen, W. L. and J. Tirado-Rives (1996). "Monte Carlo vs molecular dynamics for conformational sampling." The Journal of Physical Chemistry **100**(34): 14508-14513.
Jorgensen, W. L. and J. Tirado–Rives (2005). "Molecular modeling of organic and biomolecular systems using BOSS and MCPRO." Journal of computational chemistry **26**(16): 1689-1700.
Karplus, M. (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations." J Comput Chem **4**: 187217.
Kedzierski, L., R. L. Malby, B. J. Smith, M. A. Perugini, A. N. Hodder, T. Ilg, P. M. Colman and E. Handman (2006). "Structure of Leishmania mexicana phosphomannomutase highlights similarities with human isoforms." Journal of molecular biology **363**(1): 215-227.
Kendrew, J. C., G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff and D. C. Phillips (1958). "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." Nature **181**(4610): 662-666.
Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi (1983). "Optimization by simulated annealing." science **220**(4598): 671-680.
Kollman, P. A., I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan and W. Wang (2000). "Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models." Accounts of chemical research **33**(12): 889-897.
Kopecna, J., I. Cabeza de Vaca, N. B. P. Adams, P. A. Davison, A. A. Brindley, C. N. Hunter, V. Guallar and R. Sobotka (2015). "Porphyrin Binding to Gun4 protein, Facilitated by a Flexible Loop, Controls Metabolite Flow through the Chlorophyll Biosynthetic Pathway." Journal of Biological Chemistry.
Kozakov, D., R. Brenke, S. R. Comeau and S. Vajda (2006). "PIPER: an FFT-based protein docking program with pairwise potentials." Proteins: Structure, Function, and Bioinformatics **65**(2): 392-406.

Kozakov, D., D. R. Hall, D. Beglov, R. Brenke, S. R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili and I. C. Paschalidis (2010). "Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19." Proteins: Structure, Function, and Bioinformatics **78**(15): 3124-3130.
Krammer, A., H. Lu, B. Isralewitz, K. Schulten and V. Vogel (1999). "Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch." Proceedings of the National Academy of Sciences **96**(4): 1351-1356.
Laio, A. and M. Parrinello (2002). "Escaping free-energy minima." Proceedings of the National Academy of Sciences **99**(20): 12562-12566.
Larkin, R. M., J. M. Alonso, J. R. Ecker and J. Chory (2003). "GUN4, a regulator of chlorophyll synthesis and intracellular signaling." Science **299**(5608): 902-906.
Lasters, I., M. De Maeyer and J. Desmet (1995). "Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains." Protein engineering **8**(8): 815-822.
Leonard, G. A., T. W. Hambley, K. McAuley-Hecht, T. Brown and W. N. Hunter (1993). "Anthracycline-DNA interactions at unfavourable base-pair triplet-binding sites: structures of d(CGGCCG)/daunomycin and d(TGGCCA)/adriamycin complexes." Acta Crystallogr D Biol Crystallogr **49**(Pt 5): 458-467.
Lerman, L. (1961). "Structural considerations in the interaction of DNA and acridines." Journal of molecular biology **3**(1): 18-IN14.
Lewars, E. Computational chemistry: introduction to the theory and applications of molecular and quantum mechanics. 2010, Springer.
Li, H., W. A. Linke, A. F. Oberhauser, M. Carrion-Vazquez, J. G. Kerkvliet, H. Lu, P. E. Marszalek and J. M. Fernandez (2002). "Reverse engineering of the giant muscle protein titin." Nature **418**(6901): 998-1002.
Lindorff-Larsen, K., S. Piana, R. O. Dror and D. E. Shaw (2011). "How fast-folding proteins fold." Science **334**(6055): 517-520.
Linke, W. A., M. Kulke, H. Li, S. Fujita-Becker, C. Neagoe, D. J. Manstein, M. Gautel and J. M. Fernandez (2002). "PEVK domain of titin: an entropic spring with actin-binding properties." Journal of structural biology **137**(1): 194-205.
Lu, H. and K. Schulten (1999). "Steered molecular dynamics simulation of conformational changes of immunoglobulin domain I27 interprete atomic force microscopy observations." Chemical Physics **247**(1): 141-153.
Lu, H. and K. Schulten (1999). "Steered molecular dynamics simulations of force-induced protein domain unfolding." Proteins: Structure, Function, and Bioinformatics **35**(4): 453-463.
Lu, X. J. and W. K. Olson (2003). "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures." Nucleic acids research **31**(17): 5108-5121.
Lucas, M. F., I. C. de Vaca, R. Takahashi, J. Rubio-Martínez and V. Guallar (2014). "Atomic level rendering of DNA-drug encounter." Biophysical journal **106**(2): 421-429.
Madadkar-Sobhani, A. and V. Guallar (2013). "PELE web server: atomistic study of biomolecular systems at your fingertips." Nucleic acids research **41**(W1): W322-W328.
Mahoney, M. W. and W. L. Jorgensen (2001). "Diffusion constant of the TIP5P model of liquid water." The Journal of Chemical Physics **114**(1): 363-366.
Marquart, M., J. Walter, J. Deisenhofer, W. Bode and R. Huber (1983). "The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors." Acta Crystallographica Section B: Structural Science **39**(4): 480-490.
Marrink, S. J., A. H. De Vries and A. E. Mark (2004). "Coarse grained model for semiquantitative lipid simulations." The Journal of Physical Chemistry B **108**(2): 750-760.
Martínez-Júlvez, M., M. Medina and C. Gómez-Moreno (1999). "Ferredoxin-NADP+ reductase uses the same site for the interaction with ferredoxin and flavodoxin." JBIC Journal of Biological Inorganic Chemistry **4**(5): 568-578.
Mashiach, E., R. Nussinov and H. J. Wolfson (2010). "FiberDock: Flexible induced-fit backbone refinement in molecular docking." Proteins: Structure, Function, and Bioinformatics **78**(6): 1503-1519.

Masone, D., I. Cabeza de Vaca, C. Pons, J. F. Recio and V. Guallar (2012). "H-bond network optimization in protein–protein complexes: Are all-atom force field scores enough?" Proteins: Structure, Function, and Bioinformatics **80**(3): 818-824.
Mayoral, T., M. Martínez-Júlvez, I. Pérez-Dorado, J. Sanz-Aparicio, C. Gómez-Moreno, M. Medina and J. A. Hermoso (2005). "Structural analysis of interactions for complex formation between Ferredoxin-NADP+ reductase and its protein partners." PROTEINS: Structure, Function, and Bioinformatics **59**(3): 592-602.
Medina, M. (2009). "Structural and mechanistic aspects of flavoproteins: photosynthetic electron transfer from photosystem I to NADP+." FEBS journal **276**(15): 3942-3958.
Medina, M. and C. Gómez-Moreno (2004). "Interaction of ferredoxin–NADP+ reductase with its substrates: optimal interaction for efficient electron transfer." Photosynthesis research **79**(2): 113-131.
Messerschmidt, A., L. Prade, S. J. Kroes, J. Sanders–Loehr, R. Huber and G. W. Canters (1998). "Rack-induced metal binding vs. flexibility: Met121His azurin crystal structures at different pH." Proceedings of the National Academy of Sciences **95**(7): 3443-3448.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). "Equation of state calculations by fast computing machines." The journal of chemical physics **21**(6): 1087-1092.
Michel, J., R. D. Taylor and J. W. Essex (2006). "Efficient generalized Born models for Monte Carlo simulations." Journal of Chemical Theory and Computation **2**(3): 732-739.
Mintseris, J., K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin and Z. Weng (2005). "Protein–protein docking benchmark 2.0: an update." Proteins: Structure, Function, and Bioinformatics **60**(2): 214-216.
Miyamoto, S. and P. A. Kollman (1992). "SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models." Journal of computational chemistry **13**(8): 952-962.
Mobley, D. L., A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet and K. A. Dill (2007). "Predicting absolute ligand binding free energies to a simple model site." Journal of molecular biology **371**(4): 1118-1134.
Mongan, J., C. Simmerling, J. A. McCammon, D. A. Case and A. Onufriev (2007). "Generalized Born model with a simple, robust molecular volume correction." Journal of chemical theory and computation **3**(1): 156-169.
Morris, G. M., R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson (2009). "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility." Journal of computational chemistry **30**(16): 2785-2791.
Myers, C. and B. Chabner (1990). Cancer chemotherapy: principles and practice, Lippincott: Philadelphia**:** 256-381.
Nar, H., A. Messerschmidt, R. Huber, M. van de Kamp and G. W. Canters (1991). "Crystal structure analysis of oxidized Pseudomonas aeruginosa azurin at pH 5· 5 and pH 9· 0: A pH-induced conformational transition involves a peptide bond flip." Journal of molecular biology **221**(3): 765-772.
Nar, H., A. Messerschmidt, R. Huber, M. Van de Kamp and G. W. Canters (1992). "Crystal structure of Pseudomonas aeruginosa apo-azurin at 1.85 Å resolution." FEBS letters **306**(2): 119-124.
Nash, S. G. and J. Nocedal (1991). "A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization." SIAM Journal on Optimization **1**(3): 358-372.
Natori, R. (1954). "The role of myofibrils, sarcoplasma and sarcolemma in muscle-contraction." Jikeikai Med J **1**: 18-28.
Newman, M. and G. Barkema (1999). Monte Carlo Methods in Statistical Physics chapter 1-4, Oxford University Press: New York, USA.
Okumura, H., E. Gallicchio and R. M. Levy (2010). "Conformational populations of ligand-sized molecules by replica exchange molecular dynamics and temperature reweighting." Journal of computational chemistry **31**(7): 1357-1367.

Olsson, M. H., C. R. Søndergaard, M. Rostkowski and J. H. Jensen (2011). "PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions." Journal of Chemical Theory and Computation **7**(2): 525-537.
Onufriev, A. (2008). "Implicit solvent models in molecular dynamics simulations: A brief overview." Annual Reports in Computational Chemistry **4**: 125-137.
Onufriev, A., D. Bashford and D. A. Case (2004). "Exploring protein native states and large-scale conformational changes with a modified generalized born model." Proteins: Structure, Function & Bioinformatics **55**(2): 383-394.
Owicki, J. and H. Scheraga (1977). "Preferential sampling near solutes in Monte Carlo calculations on dilute solutions." Chemical Physics Letters **47**(3): 600-602.
Park, S., F. Khalili-Araghi, E. Tajkhorshid and K. Schulten (2003). "Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality." The Journal of chemical physics **119**(6): 3559-3566.
Pérez, A., I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton and M. Orozco (2007). "Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers." Biophysical journal **92**(11): 3817-3829.
Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten (2005). "Scalable molecular dynamics with NAMD." Journal of computational chemistry **26**(16): 1781-1802.
Pierce, B. G., Y. Hourai and Z. Weng (2011). "Accelerating protein docking in ZDOCK using an advanced 3D convolution library." PloS one **6**(9): e24657-e24657.
Pierce, M. M., C. Raman and B. T. Nall (1999). "Isothermal titration calorimetry of protein–protein interactions." Methods **19**(2): 213-221.
Ponder, J. W. (2004). "TINKER: Software tools for molecular design." Washington University School of Medicine, Saint Louis, MO **3**.
Ponder, J. W., C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht and R. A. DiStasio Jr (2010). "Current status of the AMOEBA polarizable force field." The journal of physical chemistry B **114**(8): 2549-2564.
Prinz, J.-H., H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte and F. Noé (2011). "Markov models of molecular kinetics: Generation and validation." The Journal of chemical physics **134**(17): 174105.
Qiu, D., P. S. Shenkin, F. P. Hollinger and W. C. Still (1997). "The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii." The Journal of Physical Chemistry A **101**(16): 3005-3014.
Rabi, I., S. Millman, P. Kusch and J. Zacharias (1939). "The Molecular Beam Resonance Method for Measuring Nuclear Magnetic Moments. The Magnetic Moments of Li 6 3, Li 7 3 and F 19 9." Physical review **55**(6): 526.
Ribeiro, J. M., F. Alarcon-Chaidez, I. M. Francischetti, B. J. Mans, T. N. Mather, J. G. Valenzuela and S. K. Wikel (2006). "An annotated catalog of salivary gland transcripts from Ixodes scapularis ticks." Insect biochemistry and molecular biology **36**(2): 111-129.
Rico, F., L. Gonzalez, I. Casuso, M. Puig-Vidal and S. Scheuring (2013). "High-speed force spectroscopy unfolds titin at the velocity of molecular dynamics simulations." Science **342**(6159): 741-743.
Roche, C. J., J. A. Thomson and D. M. Crothers (1994). "Site selectivity of daunomycin." Biochemistry **33**(4): 926-935.
Rosenberg, B., L. Van Camp and T. Krigas (1965). "Inhibition of cell division in Escherichia coli by electrolysis products from a platinum electrode." Nature **205**(4972): 698-699.
Rosenberg, B. and L. Vancamp (1969). "Platinum compounds: a new class of potent antitumour agents." Nature **222**: 385-386.
Ryckaert, J.-P., G. Ciccotti and H. J. Berendsen (1977). "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." Journal of Computational Physics **23**(3): 327-341.
Ryde, U., M. H. Olsson, B. O. Roos, J. O. De Kerpel and K. Pierloot (2000). "On the role of strain in blue copper proteins." JBIC Journal of Biological Inorganic Chemistry **5**(5): 565-574.

Saen-oon, S., I. C. de Vaca, D. Masone, M. Medina and V. Guallar (2015). "A theoretical multiscale treatment of protein–protein electron transfer: The ferredoxin/ferredoxin-NADP+ reductase and flavodoxin/ferredoxin-NADP+ reductase systems." Biochimica et Biophysica Acta (BBA)-Bioenergetics **1847**(12): 1530-1538.
Sastry, G. M., M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman (2013). "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments." Journal of computer-aided molecular design **27**(3): 221-234.
Scarsi, M., J. Apostolakis and A. Caflisch (1997). "Continuum electrostatic energies of macromolecules in aqueous solutions." The Journal of Physical Chemistry A **101**(43): 8098-8106.
Schaefer, M. and C. Froemmel (1990). "A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution." Journal of molecular biology **216**(4): 1045-1066.
Schaefer, M. and M. Karplus (1996). "A comprehensive analytical treatment of continuum electrostatics." The Journal of Physical Chemistry **100**(5): 1578-1599.
Schneidman-Duhovny, D., Y. Inbar, R. Nussinov and H. J. Wolfson (2005). "PatchDock and SymmDock: servers for rigid and symmetric docking." Nucleic acids research **33**(suppl 2): W363-W367.
Schrödinger, L. "New York, NY, 2010." Impact version **5**.
Segal, I. (1975). "Enzyme kinetics behaviour and analysis of rapid equilibrium and steady-state enzyme systems." A Wiley-Interscience Publication., New York, USA: 60-100.
Seibert, J. and J. Boone (1988). "X-ray scatter removal by deconvolution." Medical physics **15**(4): 567-575.
Senne, M., B. Trendelkamp-Schroer, A. S. Mey, C. Schütte and F. Noé (2012). "EMMA: A software package for Markov model building and analysis." Journal of Chemical Theory and Computation **8**(7): 2223-2238.
Setny, P. and M. Zacharias (2013). "Elastic Network Models of Nucleic Acids Flexibility." Journal of Chemical Theory and Computation **9**(12): 5460-5470.
Shaw, D. E., R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. Bank, C. Young, M. M. Deneroff, B. Batson and K. J. Bowers (2009). Millisecond-scale molecular dynamics simulations on Anton. High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on, IEEE.
Shaw, D. E., P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon and Y. Shan (2010). "Atomic-level characterization of the structural dynamics of proteins." Science **330**(6002): 341-346.
Sherman, S. E. and S. J. Lippard (1987). "Structural aspects of platinum anticancer drug interactions with DNA." Chemical Reviews **87**(5): 1153-1181.
Shih, A. Y., A. Arkhipov, P. L. Freddolino and K. Schulten (2006). "Coarse grained protein-lipid model with application to lipoprotein particles." The Journal of Physical Chemistry B **110**(8): 3674-3684.
Shih, A. Y., A. Arkhipov, P. L. Freddolino, S. G. Sligar and K. Schulten (2007). "Assembly of lipids and proteins into lipoprotein particles." The Journal of Physical Chemistry B **111**(38): 11095-11104.
Siegel, R., D. Naishadham and A. Jemal (2012). "Cancer statistics, 2012." CA: a cancer journal for clinicians **62**(1): 10-29.
Silvaggi, N. R., C. Zhang, Z. Lu, J. Dai, D. Dunaway-Mariano and K. N. Allen (2006). "The X-ray crystal structures of human alpha-phosphomannomutase 1 reveal the structural basis of congenital disorder of glycosylation type 1a." J Biol Chem **281**(21): 14918-14926.
Sitkoff, D., K. A. Sharp and B. Honig (1994). "Accurate calculation of hydration free energies using macroscopic solvent models." The Journal of Physical Chemistry **98**(7): 1978-1988.
Snow, C. D., H. Nguyen, V. S. Pande and M. Gruebele (2002). "Absolute comparison of simulated and experimental protein-folding dynamics." nature **420**(6911): 102-106.

Srinivasan, J., T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case (1998). "Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices." Journal of the American Chemical Society **120**(37): 9401-9409.
Srinivasan, J., J. Miller, P. A. Kollman and D. A. Case (1998). "Continuum solvent studies of the stability of RNA hairpin loops and helices." Journal of Biomolecular Structure and Dynamics **16**(3): 671-682.
Srinivasan, J., M. W. Trevathan, P. Beroza and D. A. Case (1999). "Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects." Theoretical Chemistry Accounts **101**(6): 426-434.
Still, W. C., A. Tempczyk, R. C. Hawley and T. Hendrickson (1990). "Semianalytical treatment of solvation for molecular mechanics and dynamics." Journal of the American Chemical Society **112**(16): 6127-6129.
Swope, W. C., H. C. Andersen, P. H. Berens and K. R. Wilson (1982). "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters." The Journal of Chemical Physics **76**(1): 637-649.
Takahara, P. M., A. C. Rosenzweig, C. A. Frederick and S. J. Lippard (1995). "Crystal structure of double-stranded DNA containing the major adduct of the anticancer drug cisplatin."
Takahashi, R., V. A. Gil and V. Guallar (2013). "Monte Carlo free ligand diffusion with Markov state model analysis and absolute binding free energy calculations." Journal of Chemical Theory and Computation **10**(1): 282-288.
Todd, R. C. and S. J. Lippard (2010). "Structure of duplex DNA containing the cisplatin 1,2-{Pt(NH3)2}2+-d(GpG) cross-link at 1.77 A resolution." J Inorg Biochem **104**(9): 902-908.
Torrie, G. M. and J. P. Valleau (1977). "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." Journal of Computational Physics **23**(2): 187-199.
Tsui, V. and D. A. Case (2000). "Molecular dynamics simulations of nucleic acids with a generalized Born solvation model." Journal of the American Chemical Society **122**(11): 2489-2498.
Ulmschneider, J. P. and W. L. Jorgensen (2003). "Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias." The Journal of chemical physics **118**(9): 4261-4271.
Vanommeslaeghe, K., E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes and I. Vorobyov (2010). "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields." Journal of computational chemistry **31**(4): 671-690.
Verdecia, M. A., R. M. Larkin, J.-L. Ferrer, R. Riek, J. Chory and J. P. Noel (2005). "Structure of the Mg-chelatase cofactor GUN4 reveals a novel hand-shaped fold for porphyrin binding." PLoS Biol **3**(5): e151.
Verdonk, M. L., J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor (2003). "Improved protein–ligand docking using GOLD." Proteins: Structure, Function, and Bioinformatics **52**(4): 609-623.
Verlet, L. (1967). "Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules." Physical review **159**(1): 98.
Vijay-Kumar, S., C. E. Bugg and W. J. Cook (1987). "Structure of ubiquitin refined at 1.8 A resolution." J Mol Biol **194**(3): 531-544.
Wang, J., R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case (2004). "Development and testing of a general amber force field." Journal of computational chemistry **25**(9): 1157-1174.
Warshel, A. and M. Levitt (1976). "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme." Journal of molecular biology **103**(2): 227-249.
Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids." Nature **171**(4356): 737-738.

Weiss, R. B. and M. C. Christian (1993). "New cisplatin analogues in development." Drugs **46**(3): 360-377.
Wilhelm, M., A. Mukherjee, B. Bouvier, K. Zakrzewska, J. T. Hynes and R. Lavery (2012). "Multistep drug intercalation: molecular dynamics and free energy studies of the binding of daunomycin to DNA." Journal of the American Chemical Society **134**(20): 8588-8596.
Xu, C., D. Tobi and I. Bahar (2003). "Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T↔ R2 transition." Journal of molecular biology **333**(1): 153-168.
Yao, S., J. P. Plastaras and L. G. Marzilli (1994). "A molecular mechanics AMBER-type force field for modeling platinum complexes of guanine derivatives." Inorganic chemistry **33**(26): 6061-6077.
Yu, Z., M. P. Jacobson, J. Josovitz, C. S. Rapp and R. A. Friesner (2004). "First-shell solvation of ion pairs: Correction of systematic errors in implicit solvent models." The Journal of Physical Chemistry B **108**(21): 6643-6654.
Zaballa, M.-E., L. A. Abriata, A. Donaire and A. J. Vila (2012). "Flexibility of the metal-binding region in apo-cupredoxins." Proceedings of the National Academy of Sciences **109**(24): 9254-9259.
Zeng, S., D. Baillargeat, H.-P. Ho and K.-T. Yong (2014). "Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications." Chemical Society Reviews **43**(10): 3426-3452.
Zhu, K., M. R. Shirts and R. A. Friesner (2007). "Improved methods for side chain and loop predictions via the protein local optimization program: Variable dielectric model for implicitly improving the treatment of polarization effects." Journal of Chemical Theory and Computation **3**(6): 2108-2119.
Zhu, K., M. R. Shirts, R. A. Friesner and M. P. Jacobson (2007). "Multiscale optimization of a truncated Newton minimization algorithm and application to proteins and protein-ligand complexes." Journal of Chemical Theory and Computation **3**(2): 640-648.