A coarse-grained view of protein-protein recognition

by

Rodney Enrique Versace Babilonia

A dissertation submitted to the Graduate Faculty in Biochemistry in partial fulfillment of the requirements for the degree of Doctor in Philosophy

The City University of New York

2012

© 2012

RODNEY ENRIQUE VERSACE BABILONIA

All Rights Reserved

This manuscript has been read and accepted for the
Graduate Faculty in Biochemistry in satisfaction of the
dissertation requirements for the degree of Doctor in Philosophy.

Dr. Marco Ceruso

_____ _____

Date Chair of Examining Committee

Dr. Edward J. Kennelly

_____ _____

Date Executive Officer

Dr. Themis Lazaridis

_____

Dr. Marilyn Gunner

_____

Dr. Michael Green

_____

Dr. Mark Kobrak

_____

Dr. K. V. Lakshmi

_____

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

A coarse-grained view of protein-protein recognition

by

Rodney Enrique Versace Babilonia

Advisor: Professor Marco Ceruso

One of the most important characteristics of proteins is their ability to specifically interact with other proteins and with other types of molecules to build supramolecular assemblies in order to perform different kinds of functions. Protein-protein interaction has been a subject of study in several sciences including biochemistry, structural biology, and computational biophysics/biology. One limitation that delays our understanding of molecular recognition is the lack of high-resolution three-dimensional structures of the protein-protein complexes. Because of this, the methods for computational prediction have gained popularity and importance but in many cases, the predicted complex is not accurate. Predicting the native three-dimensional conformation of protein-protein complexes still remains a big challenge.

Most biological processes in the cell involve a huge number of atoms and happen at time-scales that are frequently beyond the current limits of classical atomistic simulations. The ELNEDIN approach is a new and powerful coarse-grained representation with the ability to overcome size and time limits without deforming the overall shape of a protein. Previous studies

have shown that the quality of the ELNEDIN scaffold influences the ability of the modeled proteins to experience structural transitions and to associate and form a stable complex.

The main objective of this thesis is to test if the ELNEDIN approach is able to discriminate native interfaces from non-native. The Barnase / Barstar complex, the RNase / Barnase complex and the Ubiquitin / Ubiquitin ligase complex were chosen to test this hypothesis. Each individual protein model was simulated using the ELNEDIN approach and the potential of mean force for the dissociation of the complex was calculated. Our results show: 1) It is possible to obtain accurate energy-profiles using the ELNEDIN approach. 2) The shape of the free energy landscape around a protein receptor has a funnel-like shape where the bottom of the funnel is the global minimum and it starts to increase smoothly. 3) The solvent plays an important role in the shape of the free energy profile. 4) The ELNEDIN approach is able to recognize the native conformation for hydrophilic interfaces. To further support the last result, an additional complex was chosen: the Nuclease A / Nuclease inhibitor A complex.

ACKNOWLEDGEMENTS

First, I would like to thank my mentor, Dr. Marco Ceruso, for introducing me to the exciting field of theoretical and computational biophysics and for all the patience, knowledge, training and supervision in the course of this long and winding five-year road. Thanks for pushing me forward, for guiding and advising me, and for all the confidence that you showed in me. I will be eternally grateful.

A special thanks to all the members of my thesis committee for their valuable advice and unflagging availability.

Quisiera agradecer a mis padres, Lith y Víctor, por darme su infinito amor y comprensión y proveerme de las herramientas necesarias para poder salir adelante. En especial a ti, mami, por darme palabras de fuerza y aliento para seguir luchando en los momentos difíciles, te quiero mucho. Quisiera agradecer también a mi tía Nelly por todo el amor, el cariño, apoyo y comprensión que siempre me ha dado.

I would like to thank also to my wife, Setsu, for giving me hope, light, understanding, support, inspiration, strength and so much love. Thank you, honey.

Thank you to my brothers: Gary, por ser el mejor amigo del mundo y acompañarme en las buenas y en las malas desde que nos conocimos el primer día en la UPCH; Marco, por ser mi vecino, mi único compañero de laboratorio y lo mas importante por ser mi familia y aguantarme

todos estos años; vecino Lucho por alegrar mis días con tus increíbles anécdotas; Kike, por ser uno de mis mejores amigos desde que llegamos a Nueva York; and my friend Yi, for showing me his unconditional friendship in all this years of graduate student life.

Thank you to all my Japanese friends, especially to my sister Nao-chan for her beautiful friendship and for showing me that the best friends can be also found outside this galaxy.

Quisiera agradecer también a mis primos Fernando e Yrma, por darme mucha ayuda, apoyo y cariño en este lado del continente.

Thank you to my niece and nephew Teresita and Douglas for all the support and help. Especially to Douglas for taking the time and patience to correct and revise this work.

A special thanks to Dr. Maria Luisa Tasayco, for introducing me to the exciting world of Biochemistry, for giving me the chance to join her laboratory and for guiding me in the first steps of my scientific career. I am deeply grateful to you.

Thank you to all the college friends that supported and cheered me up when the going was tough: Maja, Lidia, Huan, Celia, Samar, Beicer and Allyn.

I would also like to thank the CUNY High Performance Computing Center for letting me use the high performance computational resources that permitted me to bring this project to culmination.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 Background

1.1.1 Protein-Protein Interactions

Proteins play an important role in living organisms by participating in several activities such as the catalysis of reactions, transport, the formation of building blocks of viral capsids, transmission of information from the DNA to RNA, signaling, synthesis and degradation of new molecules, and more.

One of the most important characteristics of proteins is their ability to specifically interact with other proteins and with other types of molecules to build supramolecular assemblies to perform different kinds of functions ranging from chemical catalysis to signaling and regulation (Alberts 1998). Protein-Protein recognition is the process by which the protein-protein specific interactions create functional units. These interactions have to happen at specific location or interaction sites with a specified affinity and kinetics. Aberrant protein-protein interactions have been implicated in numerous human diseases and disorders (Rual, Venkatesan et al. 2005).

Protein-protein interaction and recognition have been the subject of study of many sciences including biology, medical science, biochemistry, and biophysics (Eisenberg, Marcotte et al. 2000). One limitation that delays our complete understanding of molecular

recognition is the lack of high-resolution three-dimensional structures. The Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) contains hundreds of protein assemblies but this number is still small in comparison to the many assemblies present in a cell. Nevertheless, these hundreds of binary complexes and oligomeric proteins present in the PDB have stimulated a large number of biochemical studies by site-directed mutagenesis, supported by biophysical studies of their thermodynamic and kinetics.

Protein complexes can be classified into three groups (Janin 2009) based on the time scale of the interaction between the constituting partners:

(1) Instantaneous complexes: these complexes result from short-lived collisions that occur at every instant within the crowded space of the cell, these interactions lack biological significance but they compete with functional interactions, i.e. the crystal packing interactions in the PDB file; they are non-specific and do not play a functional role but they do produce stable assemblies, as each molecule is in contact with its neighbors (Janin 1997; Bahadur, Chakrabarti et al. 2004).

(2) Transient complexes: these complexes form when proteins associate to perform a specific task (Janin 2009). The affinities and lifetimes of transient complexes span a wide range of degrees, i.e. The Barnase / Barstar complex which has a $K_d \sim 10^{-14}$ M and a half-life of days (Schreiber and Fersht 1993).

(3) Permanent complexes: these complexes result from long-lived interactions that appear when the subunits are synthesized until they die or enter a degradation pathway (Janin 2009), i.e. the quaternary structure of some oligomers that unfold when

they are forced to unfold '*in vitro*', or when they enter the regular degradation pathway '*in vivo*'.

Protein-protein interactions may also be classified as permanent or obligatory, if the interactions between chains cannot function without each other; and transient or nonobligatory, if the interactions between chains have a molecular function also when they are in the unbound state or bound to another partner (Jones and Thornton 1996). Jones et al showed that permanent interactions are mediated by larger interfaces and that they tend to be hydrophobic, while transient interactions not only have smaller areas of contacts but in addition their nature tends to be more polar on average (Jones and Thornton 1996; Jones and Thornton 1997).

1.1.2 Definition of an interface

In order to predict protein-protein interactions, it is necessary to recognize several aspects of their association. Because proteins interact through their surfaces, it is important to gain information about their shape complementarity, the organization, the relative contribution of their physical components and their stability (Chothia and Janin 1975; Janin and Chothia 1990; Jones and Thornton 1996).

It is important to study the residues that are interacting across the two interfaces. One of the main problems here is how to define an interface, or how to define the residues that are interacting in an interface. Jones et al suggest that interface residues tend

to be more conserved than other surface residues (Jones and Thornton 1996; Jones and Thornton 1997; Ofran and Rost 2007a).

There are several definitions (Ofran and Rost 2007a; Ofran and Rost 2007b; Ofran 2009) of how to define the residues that are in the interface; the most common definition is: All residues that are solvent accessible in the unbound form but not in the bound form. But this definition depends on what is considered a 'solvent accessible' residue; it may depend on a cut-off, this cut-off could be given in terms of its absolute accessible area or in terms of the percentage of the calculated solvent accessible area. A modification of the earlier definition would then be: All residues whose accessible area was reduced upon binding. With this definition, a cutoff is unnecessary; but neither of these two definitions are able to account for interaction-dependent conformational changes.

There are many proteins that undergo conformational changes upon binding; sometimes, a buried residue that is not in the interface of the unbound form will move to the interface in the bound state. The solvent accessibility of some residues that are remote from the binding site may be affected upon conformational changes. Some definitions try to include the conformational changes: "All residues in a protein chain that are in contact with a residue in another protein chain in the bound form" (Ofran and Rost 2007b; Ofran 2009), or "Setting different distance cutoffs for each possible combination of amino acids" (Ofran and Rost 2007b; Ofran 2009). In both definitions, a cutoff has to be used in order to define when two residues are close enough to physically interact with each other. It is important to be careful when an interface needs to be defined; we have to take in

consideration possible complications, artifacts or biases that the definition may bring and try to account for them.

1.1.3 Similarities between protein folding and protein recognition

Binding is comparable with folding only when the ligands are amino acids, peptides or proteins (Tsai, Kumar et al. 1999; Tsai, Ma et al. 1999). The basic principles that determine the structure of individual proteins and protein complexes are the same:

1) Hydrophobicity plays a major role, not only in protein folding (Richards 1977; Dill 1990), but also in protein-protein interfaces (Korn and Burnett 1991; Tsai, Lin et al. 1996).

2) The concept of the energy funnel, originally postulated for protein folding, (Bryngelson and Wolynes 1989; Bryngelson, Onuchic et al. 1995; Karplus and Sali 1995; Lazaridis and Karplus 1997; Dill 1999) states that the process of protein folding goes downhill in free energy via multiple pathways instead of by following a single route. This concept has been extended to the intermolecular energy landscape of protein-protein interactions (Tsai, Kumar et al. 1999).

3) The tight packing or geometric complementarity of structural elements inside the protein is an important concept for understanding protein structure (Ponder and Richards 1987; Hubbard and Argos 1994; Jiang, Tovchigrechko et al. 2003). The same concept of complementarity is also applied to protein-protein interfaces (Keskin, Tsai et al. 2004).

1.1.4 The Association of Protein-Protein Complexes

Proteins diffuse in the cell and, upon molecular recognition events, they can form complexes. For proteins, to recognize one another and to interact is no easy task; they first have to be oriented in the correct position towards one another. A simple example to elucidate how difficult this is; it is the idea of two blind men trying to find each other in the streets of New York City (Schreiber 2002; Schreiber 2009).

Protein-protein association is believed to occur in two steps (Schreiber and Fersht 1996; Janin 1997; Selzer and Schreiber 2001; Schreiber 2002; Spaar and Helms 2005): The formation of a diffusional encounter complex, where an intermediate (A:B) is formed by diffusion, and the formation of the bound complex, where the intermediate evolves to form a final bound complex (AB) by desolvation and structural rearrangement (Figure 1.1).

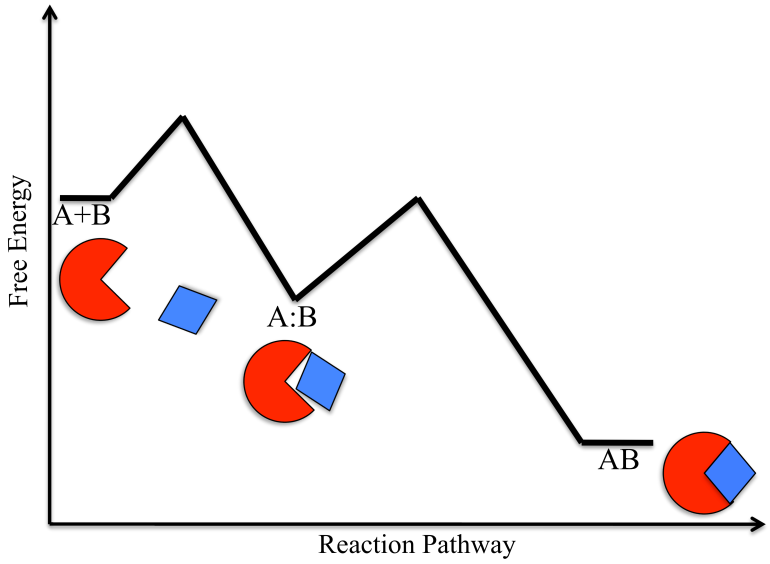

Figure 1.1: Free energy diagram describing the protein-protein binding pathway: Two proteins A (red) and B (blue) in solution will collide to form an encounter complex (A:B), which by desolvation and structural rearrangement will produce the final complex (AB).

1.1.4.1 Diffusional Encounter Complex

The formation of the encounter complex is a diffusion-controlled process. The probability of a random collision resulting in a binding event is very low (Janin 1997). The isolated ligand diffuses freely until it comes closer to the binding patch of the receptor and forms a diffusional encounter complex. This complex is not one structure; it is an ensemble of configurations able to evolve to the bound state. Association constants of about 10^6 $M^{-1}s^{-1}$ are typical of protein-protein complexes that bind without strong electrostatic interaction. The presence of attractive electrostatic forces can lead to higher rates very close to 10^{10} $M^{-1}s^{-1}$. Electrostatic attraction and repulsion are long-range forces that may guide the association process; the electrostatic attraction is the factor that contributes more to the rate of association (Camacho, Weng et al. 1999; Gabdoulline and Wade 2001; Gabdoulline and Wade 2002; Pachov, Gabdoulline et al. 2009).

1.1.4.2 Bound Complex

After formation of the diffusional encounter complex, the two proteins must adjust their positions to form the final bound complex. They may undergo translations, rotations and conformational changes. The intermolecular forces present to hold together the biomolecules are short-range noncovalent interactions such as salt-bridges, hydrogen bonds, van der Waals interactions and hydrophobic forces (Janin 1997; Gabdoulline and Wade 2001; Selzer and Schreiber 2001; Gabdoulline and Wade 2002).

One of the main problems in determining the pathway two proteins use to become a complex is the determination of the rate-determining step of the protein-protein association process: some authors (Janin 1997; Selzer and Schreiber 2001; Spaar and Helms 2005) propose that the rate-determining step is the transition from the encounter complex to the final complex, A:B → AB, (Figure 3.1 A). Another study (Selzer and Schreiber 2001) shows the rate determining step is the formation of the encounter complex from the dissociated proteins, A+B → A:B. Protein-protein association studies for the Barnase / Barstar complex showed that the formation of the final complex is a downhill process (Spaar and Helms 2005; Spaar, Dammer et al. 2006; Hoefling and Gottschalk 2010; Wang, Siu et al. 2010).

1.1.5 Protein-Protein Docking

If the structures of the unbound form of two proteins exist in the Protein Data Bank, but the 3D structure of the complex was not yet obtained, we have to rely on computational predictions (Ofran 2009). The field of modeling the structure of complexes is called 'Docking'. It is an *in silico* method that tries to predict a protein-protein complex of two proteins based on the coordinates of the unbound partners (Hwang, Pierce et al. 2009), but sometimes these predictions are not accurate or provide us with a false positive. For that reason the problem of finding the correct native three-dimensional configuration is still a big challenge.

Docking can be classified into two categories: bound docking, when the structures of the complex are used as input, and unbound docking, when the individual unbound structures are used as input. The first is used only for verification and testing purposes, while the latter is where all the current docking research is focused.

Protein docking involves searching regions of shape complementarity but conformational changes have to be allowed upon binding. Overlapping and flexibility must be taken into account. Docking is usually divided in two stages or phases. In the first stage the ligand and the receptor are treated as rigid bodies with a soft scoring function that allows overlap. In this stage several thousand complexes are commonly generated. In the second, or refinement stage, the structures obtained in the initial stage suffer small flexible motions of the side chains and/or backbone, followed by a detailed rescoring which will filter out the false positives produced by the initial state. (Figure 3.4) (Chen, Li et al. 2003).

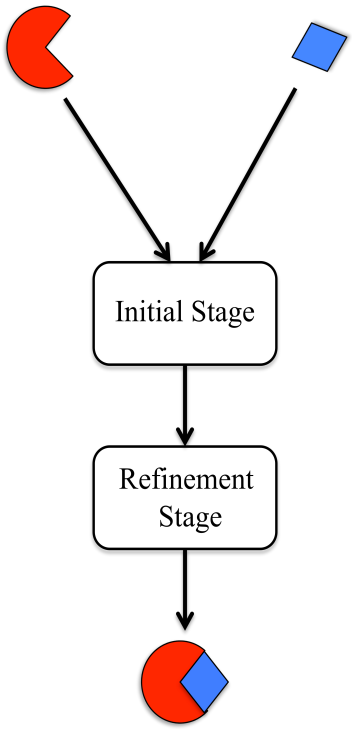

Figure 1.2: The general protein-protein docking process: Two stage approach to protein docking of two proteins: receptor (red) and ligand (blue).

1.2.- Can the ELNEDIN approach be used as a tool to discriminate non-native from native interfaces?

Most biological processes in the cell happen at time-scales and involve a huge number of atoms that are frequently beyond the current limits of the classical atomistic simulations. The ELNEDIN approach is a new and powerful coarse-grained representation with the ability to overcome size and time limits without deforming the overall shape of a protein. Previous studies (Periole, Cavalli et al. 2009) have shown that the ELNEDIN approach is able to simultaneously reproduce the local and global deformations of a protein allowing stable microsecond time scale molecular dynamics simulations. The quality of the ELNEDIN scaffold influences the ability of the modeled proteins to experience structural transitions and to associate and form a stable complex.

The main objective of this work is to test if the ELNEDIN approach is able to identify the non-native interfaces and recognize only the native or native-like configuration. To this end, three protein-protein complexes were chosen:

1) The Barnase / Barstar complex (PDB I.D.: 1BRS), the most commonly used binary complex in protein literature, which presents a high affinity value ($K_d = 2.0x10^{-13}$ / ΔG = -17.3 kcal/mol).

2) The RNase / Barstar complex (PDB I.D.: 1AY7), another Barstar containing complex but of lower affinity ($K_d = 2.0x10^{-10}$ / ΔG = -13.2 kcal/mol); than the previous one.

3) The Ubiquitin / Ubiquitin Ligase complex (PDB I.D.: 2OOB) a low affinity complex ($K_d = 6.0x10^{-5}$ / ΔG = -5.7 kcal/mol).

Each individual protein model was simulated using the ELNEDIN approach emplying a rigid elastic network, and the potential of mean force was calculated and analyzed. To validate the quality of the free energy profiles obtained by using the ELNEDIN representation, atomistic approaches have been performed on those three complexes. To further support the result of the main hypothesis, an additional complex, which undergoes large conformational changes upon complexation, was chosen: the Nuclease A / Nuclease inhibitor A complex.

2

Methods

2.1 Theory

2.1.1 Molecular Dynamics

2.1.1.1 Molecular Dynamics Simulations

Molecular dynamics is a computer simulation technique that follows the time evolution of a set of interacting atoms by solving Newton's equation of motion (equation 2.1), allowing us to study the behavior of several molecular systems. If we consider a system of N particles, each particle has a starting position r and velocity v:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = \overrightarrow{F_i} \qquad 2.1$$

where m_i is the atom mass, and $\overrightarrow{F_i}$ is the force on each atom, considering the interaction with other atoms. The force $\overrightarrow{F_i}$ also can be defined as the negative derivative of the potential energy function (V):

$$\overrightarrow{F_i} = -\frac{\partial V}{\partial r_i} \qquad 2.2$$

Molecular dynamic simulations require special software to integrate the equation of motion of the inter-atomic interactions, generating a trajectory output file where the coordinates are written as a function of time.

The most commonly used algorithm to integrate the equation of motion is called the leap-frog algorithm (equation 2.3 and 2.4) (Hockney, Goel et al. 1974), a variation of the third order and time reversible Verlet algorithm (Verlet 1967). The leap-frog algorithm uses the position r at time t and velocity v at time $t - \frac{\Delta t}{2}$. The velocities and positions are updated using the forces $F(t)$ determined by equation 2.2 at time t:

$$v\left(t + \tfrac{\Delta t}{2}\right) = v\left(t - \tfrac{\Delta t}{2}\right) + \frac{F(t)}{m}\Delta t \qquad 2.3$$

$$r\left(t + \Delta t\right) = r(t) + v\left(t + \tfrac{\Delta t}{2}\right)\Delta t \qquad 2.4$$

The force field describes the potential energy function V. It contains a set of mathematical functions describing the different contributions to the total potential energy. Commonly, force fields are classified in three main groups:

a) 'All-atom' force fields, which provide the parameters for all the atoms in the system, including hydrogens, e. g. CHARMM (MacKerell 1998), OPLS (Jorgensen and Tirado-Rives 1988) and AMBER (Cornell, Cieplak et al. 1995; Hornak, Abel et al. 2006) force field;

b) ‘United-atom’ force fields, which treat methyl and methylene groups as a single interaction center, e.g. GROMOS96 force field (Van Gunsteren, Billeter et al. 1996); and,

c) ‘Coarse-grained’ force-fields, a simplified representation obtained by merging several atoms in one bead, e.g. Martini force field (Marrink, deVries et al. 2004).

Each force field has a specific way to decompose the contributing terms for the total potential energy, but the general form can be written as:

$$V = V_{bonded} + V_{nonbonded} \qquad 2.5$$

The bonded energy terms define the covalent energy of the system, and it is given by the sum of the terms (equation 2.6) corresponding to the bond (equation 2.7), angle (equation 2.8), dihedrals (or torsion) and improper dihedrals (out of plane distortions) (equation 2.9):

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} \qquad 2.6$$

$$V_{bond} = \sum_{bonds} \frac{1}{2} k_r \left(r - r_0 \right)^2 \qquad 2.7$$

$$V_{angle} = \sum_{angles} \frac{1}{2} k_\theta \left(\theta - \theta_0 \right)^2 \qquad 2.8$$

$$V_{dihedral} = \sum_{dihedrals} \frac{1}{2} V_n \left[1 + \cos\left(n\theta - \delta \right) \right] \qquad 2.9$$

The bond and angle potentials (equation 2.7 and 2.8) are usually treated as harmonic oscillators.

The improper dihedrals (equation 2.10) are defined to maintain the planarity or to avoid molecules flipping over to their mirror image. This type of dihedral is treated also as a harmonic potential.

$$V_{improper} = \frac{1}{2} k_{\xi} (\xi_{ijkl} - \xi_0)^2 \qquad 2.10$$

The non-bonded energy terms define the interaction between atoms that are not connected by covalent bonds (equation 2.11). It is given by long-range electrostatics, computed using the Coulomb's law (equation 2.12), and the van der Waals forces, calculated with the Lennard-Jones potential which describes the long-range dispersions and the short-range repulsive interactions (equation 2.13). Both energies are pairwise additive because each interaction is treated separately and is independent from each other, reducing in this way the computational time.

$$V_{non-bonded} = V_{Electrostatic} + V_{LJ} \qquad 2.11$$

$$V_{Electrostatic} = \frac{1}{4\pi\varepsilon_0 \varepsilon} \sum_{ij\ pairs} \frac{q_i q_j}{r_{ij}} \qquad 2.12$$

$$V_{LJ} = \sum_{ij\ pairs} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \qquad 2.13$$

Molecular dynamics is limited by the memory and the speed of the computer; the size of the system is the main factor in determining the computational cost due to the non-bonded energy calculations. For that reason it is important to define the boundaries of the system. But working with finite systems produces some artifacts in the boundaries;

one way to avoid this problem is by using Periodic Boundary Conditions (PBC) (Bekker, Dijkstra et al. 1995), this consists of placing the molecules inside a box of penetrable walls, but with the same number of molecules at any given time. In other words, if one molecule leaves the box through one side, it must re-enter with the same velocity through the opposite side (Alder and Wainwright 1959).

2.1.1.2 Atomistic Models: The AMBER99SB Force Field

AMBER (Assisted Model Building with Energy Refinement) is a family of force fields for molecular dynamics of bio-molecules developed by Peter Kollman's group at the University of California, San Francisco. The most widely used version, known as ffamber94 (Cornell, Cieplak et al. 1995), was inspired by the OPLS (Optimized Potential for liquid simulations) force field (Jorgensen and Tirado-Rives 1988; Jorgensen, Maxwell et al. 1996). The bonded and nonbonded parameters are described by equations 2.6 – 2.13.

Because ffamber94 strongly favors helical conformations, several updates were done. The third generation of these updates, ffamber99 (Wang, Cieplak et al. 2000; Sorin and Pande 2005), is most commonly used for protein and nucleic acids. It conserved most of the main features of ffamber94. But still, it has the tendency to favor helical conformations. The ffamber94 and ffamber99 versions failed to describe accurately two sets of backbone φ/ψ dihedral parameters. Looking to achieve a better balance of the secondary structure types and also to improve the φ/ψ dihedral parameters terms existing in the ffamber99 energy function, ffamber99SB was released (Hornak, Abel et al. 2006).

2.1.1.3 Coarse Grained Models: The Martini Force Field

Atomistic simulations of biological systems that are involved in key biological processes such as protein-protein interaction, protein aggregation, ligand binding, protein-membrane simulations, etc., are restricted to small length and time scales. The huge number of atoms involved in these processes makes them difficult to study using atomistic approaches. To overcome the time scale and size limitations without sacrificing the resolution of the biological system and to reproduce accurately the structure, dynamical properties, and the several transient intermolecular interactions is a challenging task. Several approaches were developed, such as the use of algorithms to enhance the conformational sampling, but these kinds of approaches are limited to atomistic dynamic simulations (Tozzini 2005; Periole, Cavalli et al. 2009).

Another approach that consists of using a simplified model to reduce the number of degrees of freedom, thus increasing the time-scale of the simulation and the size of the system, is the development of coarse-grained (CG) models capable of yielding an accurate description of the free energy surface. Coarse-grained models have been developed to achieve longer time simulations of bigger systems, such as lipid membranes and proteins. In this kind of approach, small groups of atoms are merged, forming beads. Each bead is based on a united atom representation and is represented by one or more interaction sites (Marrink, deVries et al. 2004). This model was originally designed for lipid and surfactant systems.

According to the number of interacting centers present in a bio-molecule (e.g. amino acids), we can group the model from the coarsest (one bead) to the finest (four to

six beads). But the coarser the model, the harder is the parameterization of the force field, and there is a loss in accuracy and transferability (Tozzini 2005). These CG models are parameterized following a structural-based or a physics-based approach. The Martini force field is a physics-based parameterization of a large library of building blocks against experimental thermodynamic data (Periole, Cavalli et al. 2009).

The original Martini force field for lipids was modified to make it suitable for bio-molecular simulations (Marrink, Risselada et al. 2007). It is based on a 4 to 1 parameterization (four heavy atoms represented by a single interaction center). It has been defined by four types of CG beads: polar (P), non-polar (N), apolar (C), and charged (Q). N and Q were subdivided based on the ability to form hydrogen bonds in "0" (no formation), "d" and "a" (hydrogen bond donor and acceptor respectively), "da" (hydrogen bond donor and acceptor).

The bonded parameters are described by the following sets of potentials:

$$V_{bond} = \frac{1}{2} k_{bond} (r - r_{bond})^2 \qquad 2.14$$

Where the equilibrium distance r_{bond} is equal to the predetermined value of 0.47nm and the force constant k_{bond} equal to $1250 kJ mol^{-1} nm^{-2}$.

$$V_{angles} = \frac{1}{2} k_{angle} [\cos(\theta) - \cos(\theta_0)]^2 \qquad 2.15$$

Where k_{angle} and θ_0 can take different values according of the nature of the bond, e.g.: for aliphatic chains $k_{angle} = 25 kJ mol^{-1}$ and $\theta_0 = 180°$, for cis double bonds $k_{angle} = 45 kJ mol^{-1}$ and $\theta_0 = 120°$.

The type of the backbone particle depends on the secondary structure of the protein, e.g.: the backbone particle of Ala is C5 if it is part of a helix, N0 if it is in a helix at N- or C-terminus or β–strand, or P4 when it is part of a coil.

$$V_{improper} = k_{improper} (\theta - \theta_{improper})^2 \qquad 2.16$$

The improper dihedral angle (equation 2.16) is used to maintain the planarity of planar groups like rings. To model the ring structure, a new CG bead 'S' was introduced with a 2 or 3 to 1 mapping.

The Coulomb law and a shifted LJ 12-6 potential describe the non-bonded interactions:

$$V_{electrostatic} = \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r} \qquad 2.17$$

where the relative dielectric constant $\varepsilon_r = 15$.

$$V_{LJ} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \qquad 2.18$$

A σ = 0.47nm is assumed for each interacting pair, with exception of two particular cases. There are established ten levels of LJ interactions for the value of ε (Table 2.1)

Table 2.1: Levels of LJ interactions: List of the 10 predetermined values of the LJ parameters that indicates the well depth.

Level of interaction	ε
Supra attractive	5.6
Attractive	5.0
Almost attractive	4.5
Semi attractive	4.0
Intermediate	3.5
Almost intermediate	3.1
Semi repulsive	2.7
Almost repulsive	2.3
Repulsive	2.0
Super repulsive	2.0 (σ=0.62)

Four water molecules are represented as a single type P_4. An antifreeze water bead particle big-P_4 (BP_4) was introduced in order to avoid the higher freezing temperature of the CG water (P_4). The substitution of the 10% of the system P_4 waters with BP_4 might prevent the freezing and basically maintain the water fluid for a longer time.

The latest version of the MARTINI force field (MARTINI 2.1) was released in 2008 (Monticelli, Kandasamy et al. 2008) and represents the extension of the CG model to polypeptides and proteins.

2.1.1.4 The ELNEDIN representation

One way to solve these problems in the coarsest model is through the use of an elastic network model, where the system is represented by a network of interacting backbone beads, usually one bead per amino acid (Cα), linked by elastic springs or harmonic potentials (Figure 2.1).

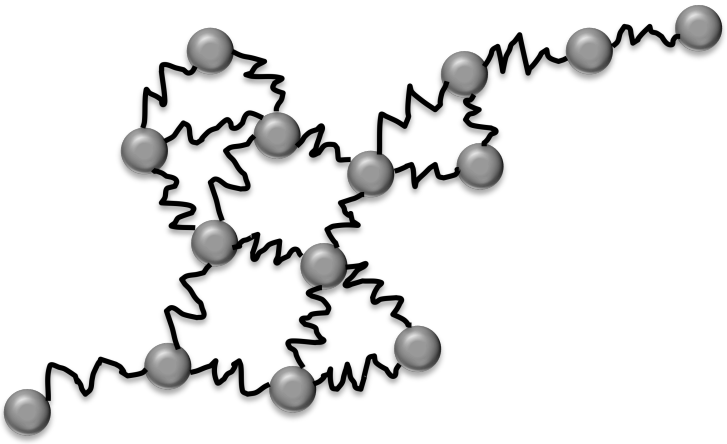

Figure 2.1: Cartoon representation of an elastic network: Two spheres are connected by a spring with a force constant of K_{SPRING}, if the spheres are less than R_C Å.

Elastic Network (EN) models are considered structural-based force fields. They were proposed by Tirion (Tirion 1996) to eliminate the costly computational

minimization procedure in classical atomistic normal mode analysis by using the native structure as the minimum of the free energy. This approach introduced an intrinsic bias toward the initial experimental structure. EN models are described as a network of point masses attached via springs with a force constant (K_{SPRING}) when the distance between them is less than a cut-off distance (R_C). These two parameters characterize the rigidity and the extent of the network.

Previously, our laboratory has investigated the possibility of combining a structure-based model (an elastic network (Tirion 1996)) and a physics-based coarse-grained model (the Martini force field (Marrink, Risselada et al. 2007)) into a unique representation called ELNEDIN (Periole, Cavalli et al. 2009). In an ELNEDIN model the elastic network acts as a structural scaffold and the Martini force field directs the intermolecular interactions. The results showed that the quality of the scaffold influences the ability of the modeled proteins to experience structural transitions and to associate and form a stable complex.

The system is represented by a network of interacting backbone beads, usually one bead per amino acid (Cα), linked by elastic springs or harmonic potentials. The energy between two beads i and j is given by:

$$E_{ij} = \frac{1}{2} K_{SPRING} \left(r_{ij} - r_{ij}^{0} \right)^2 \qquad (3.1)$$

Where K_{SPRING} is the force constant, r_{ij} is the distance between the beads, and r_{ij}^{0} is the reference distance from the experimental model. Beads are considered bound to each other if their distance is less than a predefined cut-off distance, R_C and if they are

separated by at least two positions in the protein sequence (Figure 2.1). Values of K_{SPRING} and R_C may vary depending on the kind of desired experiment to perform. These variations produce more or less stiff network representations (Figure 2.2).

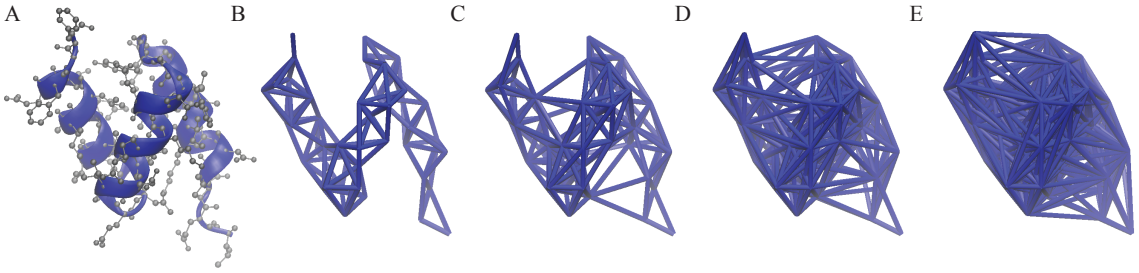

Figure 2.2: ELNEDIN scaffolds of E3 Ubiquitin ligase produced by different R_C: A) all atoms representation. B) R_C = 0.6 nm. C) R_C = 0.8 nm. D) R_C = 1.0 nm. E) R_C = 1.2 nm.

The current version of the ELNEDIN representation is based on the MARTINI 2.1 force field. The main modification applied to the ELNEDIN approach is that the position of the backbone beads were placed on the Cα instead of the center of mass of the atoms of the backbone (N, Cα, C, O). This modification was done in order to make the ELNEDIN representation independent from the secondary structure of the system. The treatment of the non-bonded interaction was not modified with respect to the MARTINI force field, but the bonded interactions between backbone beads, backbone and side chains, and between side chains were reparametrized (Ceruso, Periole et al. 2004).

2.1.2 Potential of Mean Force and Umbrella Sampling

The potential of mean force (PMF) (Kirkwood 1935) along a reaction coordinate is the calculation of the free energy changes as a function of a coordinate of the system. The choice of this reaction coordinate or degree of freedom (ξ) is very important, and should be related to the question to be answered. It can be a simple function of the Cartesian coordinates such as a distance (e.g. the distance between the center of mass of two molecules: Barnase and Barstar), an angle or dihedral (e.g. the Φ and Ψ angles of Alanine Dipeptide), or it can have more complicated forms, depending upon the system or process of study.

Simple molecular dynamic simulations are limited and are not a good choice for calculating the PMF within the available computational resources and time, because the sampling will be limited only to the local minimum due to its inability to cross the energetic barriers. To overcome this issue, several methods were designed to improve the sampling in those 'forbidden' region, making possible to calculate the PMF from a MD trajectory.

Umbrella sampling (Torrie and Valleau 1977) is a technique to enhance sampling in the vicinity of a chosen value of a reaction coordinate ξ by the insertion of a positive biasing window potential $W_{(\xi)}$ to the potential energy $V_{(r)}$.

$$V' = V_{(r)} + W_{(\xi)} \qquad 2.19$$

The biasing potential restricts the variations of the reaction coordinate within a small interval around a chosen value. Because the sampling was done only for this small region or window, to obtain the PMF of the complete region a series of calculations must be performed to populate the different but overlapping regions. For each window i is a harmonic potential of the

form of equation 2.20 which will be applied:

$$W_{(\xi)i} = \frac{1}{2} k_W (\xi - \xi_0)^2 \qquad 2.20$$

Where k_W is the umbrella force constant, ξ is the actual distance along the reaction coordinate, and ξ_0 is the center of the window i. To reconstruct the full distribution function from the separate distributions of each window, the trajectories of each window are used to calculate a set of biased distribution function, in a form of overlapping histograms, for the complete range of the reaction coordinate (Roux 1995).

From the biased distribution functions, it is possible to calculate the unbiased distribution functions. This can be done using the weighted histogram analysis method (WHAM) (Kumar, Bouzida et al. 1992). The WHAM algorithm is based on the maximum overlap method to calculate free energy (Bennet 1976) to construct an estimate of the unbiased distribution function as a weighted sum over the data obtained from all the trajectories, and determine the functional forms of the weight factors that produce the lower statistical error.

2.1.3 Hierarchical Clustering

Clustering is a statistical technique that allows for classifying data in an evolutionary context (Rizzi, Mahata et al.). In other words, it allows one to group or segment a set of data into subsets or 'clusters', in a way that elements inside each cluster are more closely related to one another than elements assigned to a different cluster (Hastie, Tibshirani et al. 2009).

Hierarchical clustering is divided into agglomerative methods (Figure 2.3) which merges clusters iteratively, and divisive methods (Figure 2.3), which group all objects into one cluster and subdivide them into smaller pieces.

The basic methods of agglomerative hierarchical clustering (Johnson 1967) for an NxN distance (or similarity) matrix consists of:

1. Starting by assigning each cluster item to a cluster.
2. Finding the closest pair of clusters (the most similar) and merge them into a single cluster.
3. Computing distance between the new and each of the old clusters using any of the several algorithms: single linkage, complete linkage, average linkage, median distance linkage (D'andrade 1978), etc. In our analysis we used the average linkage method, where the distance between one cluster and another is equal to the average of distances between all pairs of elements, where each pair is made up of one element of each cluster (Borgatti 1994).
4. Repeating steps 2 and 3 until all items are clustered into a single cluster.

Hierarchical clustering may be represented by a two dimensional diagram called dendrogram (Figure 2.3).

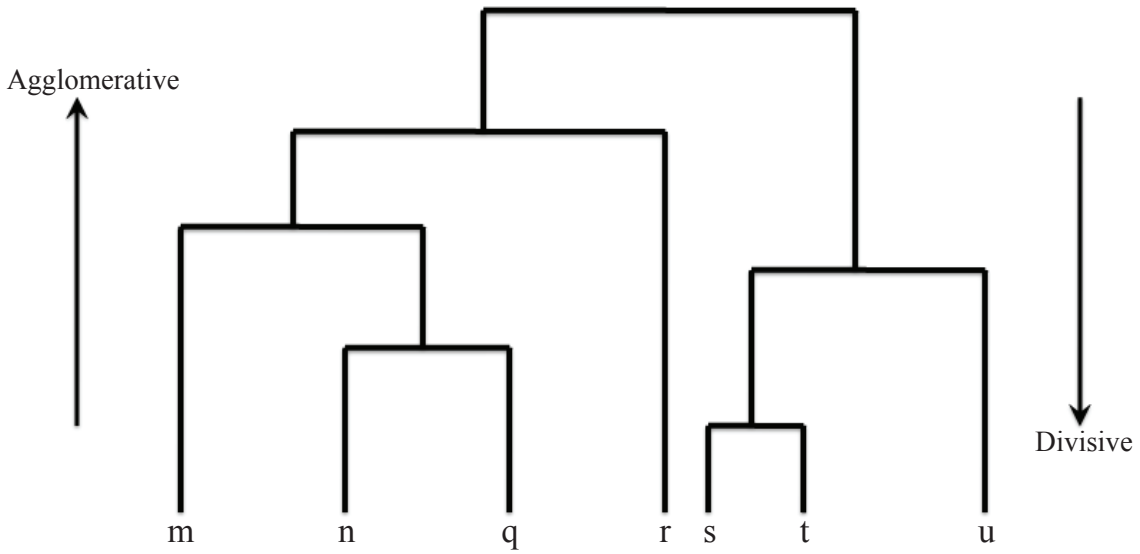

Figure 2.3: Schematic representation of a dendrogram: The agglomerative and divisive methods are shown.

2.2 Protocols and Parameters

2.2.1 Studied Systems

2.2.1.1 The Database

With the goal of testing the effectiveness of our methodology, 44 complexes were chosen. 43 complexes were chosen from the rigid body set of proteins of the Protein Docking Database Benchmark 4.0 (Hwang, Vreven et al. 2010). The 44th complex is not in the database, it was chosen because it is one of the most common studied systems in the literature of the last few years: the Barnase-Barstar complex.

There were two conditions to select the 43 complexes: Each complex should be composed by two chains (a receptor and a ligand) and each chain should not have missing residues. Missing atoms were added with the bioinformatics tool Swiss PDB viewer (Guex and Peitsch 1997). The Protein-Protein Docking Benchmark classifies the complexes in three levels of difficulty based on if the interface residues of the ligand and/or the receptor may undergo conformational changes upon complex formation: Easy or Rigid Body, Medium and Difficult (Chen, Mintseris et al. 2003; Hwang, Pierce et al. 2008; Hwang, Vreven et al. 2010). Of the 43 complexes, only 2O3B belongs to the difficult level, the other 42 were taken from the easy set (Table 2.2).

Table 2.2: The Benchmark: List of the system that were used to calculate the BSA and SC values.

Complex	Chain A	Chain B
1BRS	Barnase	Barstar
1AY7	RNase SA	Barstar
2OOB	E3 Ubiquitin ligase	Ubiquitin
1Z5Y	N-term of DsbD	E. coli CCMG protein
1KAC	Adenovirus Fiber knob protein	Adenovirus receptor
1UDI	Uracyl-DNA glycosylase	Glycosylase inhibitor
1YVB	Falcipain 2	Cystatin
2B42	Xylanase	Xylanase inhibitor
2SNI	Subtilisin	Chymotrypsin inhibitor 2
1OPH	α-1-antitrypsin	trypsinogen
1AK4	Cyclophilin	HIV capsid
1B6C	FKBP binding protein	TGFβ receptor
1BVN	α-amylase	Tendamistat
1CGI	Bovine chymotrypsinogen	PSTI
1CLV	α-amylase	α-amylase inhibitor
1D6R	Bovine trypsin	Bowman-Birk inhibitor
1EAW	Matriptase	BPTI
1FC2	Staphylococcus Protein A	Human Fc fragment
1FFW	Chemotaxis protein CheY	Chemotaxis protein CheA
1FLE	Elastase	Elafin
1GCQ	GRB2 C-ter SH3 domain	Vav N-ter SH3 domain
1GHQ	Complement C3	Epstein-Barr virus receptor CR2
1GPW	HISF protein	Amidotransferase HISH
1HE1	Rac GTPase GNP	Pseudomonas toxin GAP dom
1KXQ	Camel VHH	Pancreatic α-amylase
1PPE	Bovine trypsin	CMTI-1 squash inhibitor
1PVH	IL6 receptor β chain D2-D3 domains	Leukemia inhibitory factor

1QA9	CD2	CD58
1R0R	Subtilisin Carlsberg	OMTKY
1SBB	T-cell receptor β	Staphylococcus enterotoxin B
1TMQ	α-amylase 5HP	RAGI inhibitor
1XD3	UCH-L3	Ubiquitin
1Z0K	Rab4A GTPase GNP	RAB4 binding domain of Rabenosyn
2BTF	Actin ATP	Profilin
2I25	Shark single domain antigen receptor	Lysozyme
2J0T	MMP1 intersitial collagenase	Metalloproteinase inhibitor 1
2OUL	Falcipain 2	Chagasin
2PCC	Cyt C peroxidase HEM	Cytochrome C
2SIC	Subtilisin	Streptomyces subtilisin inhibitor
2UUY	Trypsin	Tryptase inhibitor from tick
2VDB	Serum albumin	Peptostreptococcal albumin-biding protein
3D5S	Complement C3d fragment	Fibrinogen-binding protein C-ter domain
3SGQ	Streptogrisin B	Ovomucoid inhibitor third domain
2O3B	NucA nuclease	NuiA nuclease inhibitor

The accuracy of the prediction was measured using the root mean square deviation (RMSD) (equation 2.21) between the ligand of the predicted conformation (p) and the ligand of the crystal structure of the complex (n), the one with the lowest RMSD value should correspond to the 'native-like' conformation. The receptors of the predicted and native complex were previously structurally aligned. And the RMSD value was calculated on the Cα atoms of the ligands.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\{\left[x_p(i) - x_n(i)\right]^2 + \left[y_p(i) - y_n(i)\right]^2 + \left[z_p(i) - z_n(i)\right]^2\right\}} \qquad 2.\ 21$$

Where N is the total number of atoms and x,y,z are the Cartesian coordinates of the atoms of the predicted and native protein ligands.

2.2.1.2 The Chosen Systems

From the 44 complexes, 4 protein-protein complexes have been chosen to answer the main question of this work; each protein has high resolution structures available in the Protein Data Bank (Berman, Westbrook et al. 2000) and for which dissociation constants (Kd) were measured by experimental biophysical methods. The respective Kd of these three proteins could be classified as high, medium and low affinity (Kastritis, Moal et al. 2011) (Table 2.3).

Table 2.3: Studied Systems: List of chosen complexes.

Complex	Kd (M)	ΔG_{exp} (kcal/mol)
1BRS[1]	2.0E-13	-17.3[2]
1AY7[3]	2.0E-10	-13.2[4]
2OOB[5]	6.0E-5	-5.7[6]
2O3B[7]	3.2E-12	-15.7[7]

(1) (Buckle, Schreiber et al. 1994). (2) (Hartley 1993). (3) (Sevcik, Urbanikova et al. 1998). (4) (Hartley, Both et al. 1996). (5) (Peschard, Kozlov et al. 2007). (6) (Kozlov, Nguyen et al. 2007; Kozlov, Peschard et al. 2007). (7) (Ghosh, Meiss et al. 2007).

2.2.1.2.1 The Barnase / Barstar complex (1BRS)

The Barnase / Barstar complex (Hartley 1993; Buckle, Schreiber et al. 1994) became a very well-studied target for computational and experimental studies because it is a small system with fast associations rates and one of the most stable complexes known (Figure 2.4). Its association is electrostatically facilitated and it is considered one of the most strongest between proteins (Spaar, Dammer et al. 2006).

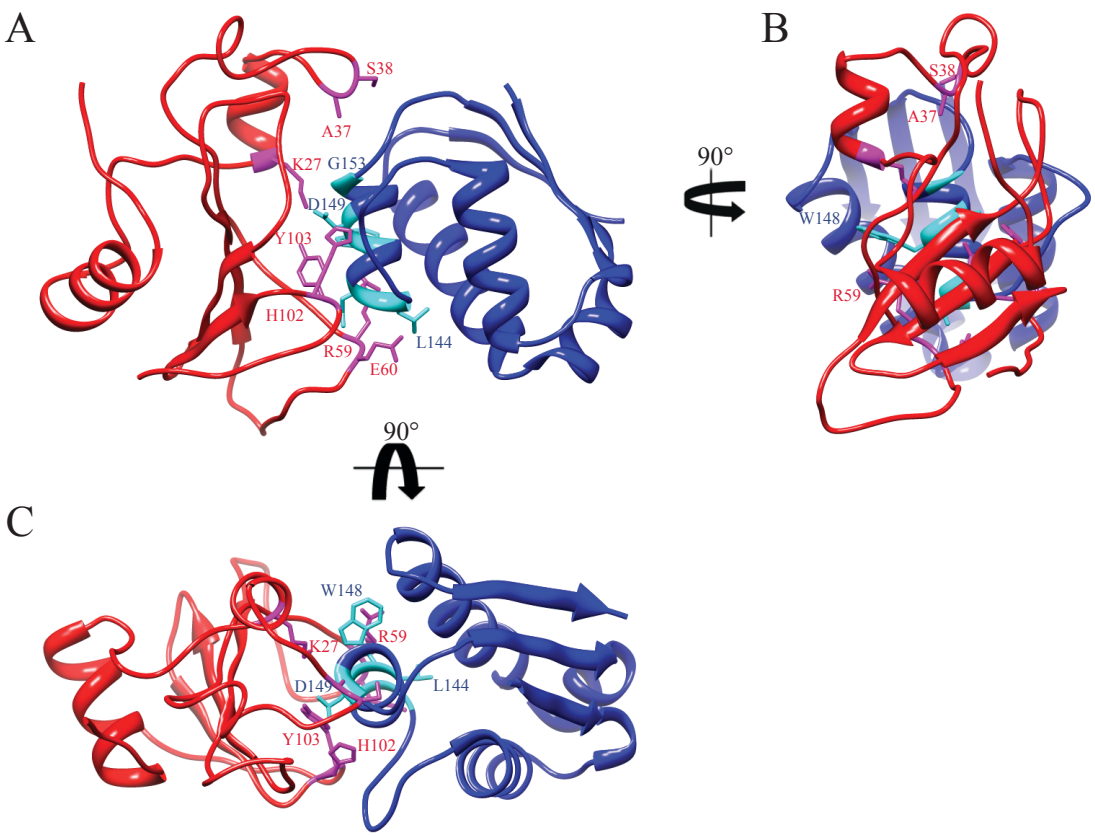

Figure 2.4: The Barnase / Barstar complex: Cartoon representation of the wild type Barnase (red) in complex with Barstar (blue). Barnase and Barstar interface residues depicted in magenta and cyan respectively. A) Front view. B) Side view. C) Top view.

Barstar is the natural specific intracellular inhibitor of Barnase; the same organism, *Bacillus amyloliquefaciens*, produces both of them. Barnase is a 110-residues RNAse that degrades RNA snippets. Its interface is very polar and contains several positive charged residues such as lysine and arginine. Barstar has only 89 residues and it bears a strong negative binding interface (Figure 2.5). The way that Barstar inhibits Barnase is by sterically blocking the active site of Barnase with a helix and an adjacent loop segment (Figure 2.4). This inhibition is a self defense mechanism of the hosting organism to protect itself, since expression of Barnase is lethal in the absence of Barstar. Barnase needs to complex with Barstar from the moment of expression until secretion;

for this reason the association has to be very rapid and strong to prevent cell death. There is a high degree of shape and charge complementarity of the interacting surfaces that favorites the association and binding rate, making the Barnase / Barstar complex not only one of the best-studied systems, but also a very useful complex in different fields such as biochemistry, biophysics and bioengineering (Strittmatter, Janssens et al. 1995).

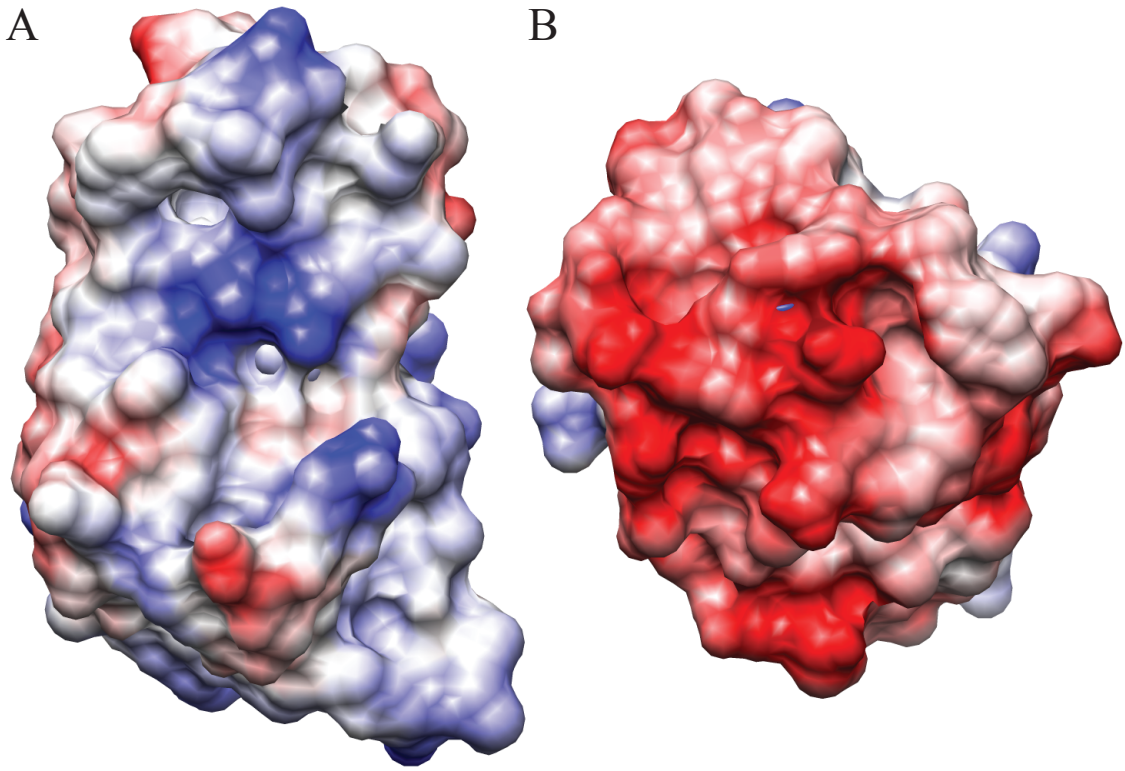

Figure 2.5: Electrostatics map of the Barnase / Barstar interface: Each member of the complex was rotated 90° around the vertical axis. The surface color code goes from positive (blue) to negative (red). A) Barnase. B) Barstar.

2.2.1.2.2 The RNase / Barstar complex (1AY7)

Despite being produced by different organisms, Barstar also inhibits RNase SA in the same way as it does to Barnase, by sterically blocking the active site. RNase SA has been isolated from *Streptomyces aureofaciens* bacteria and is a single-chain protein with 96 residues (Figure 2.6). *Streptomyces aureofaciens* also produces its own specific ribonuclease inhibitor but as of today, the isolation remains extremely difficult.

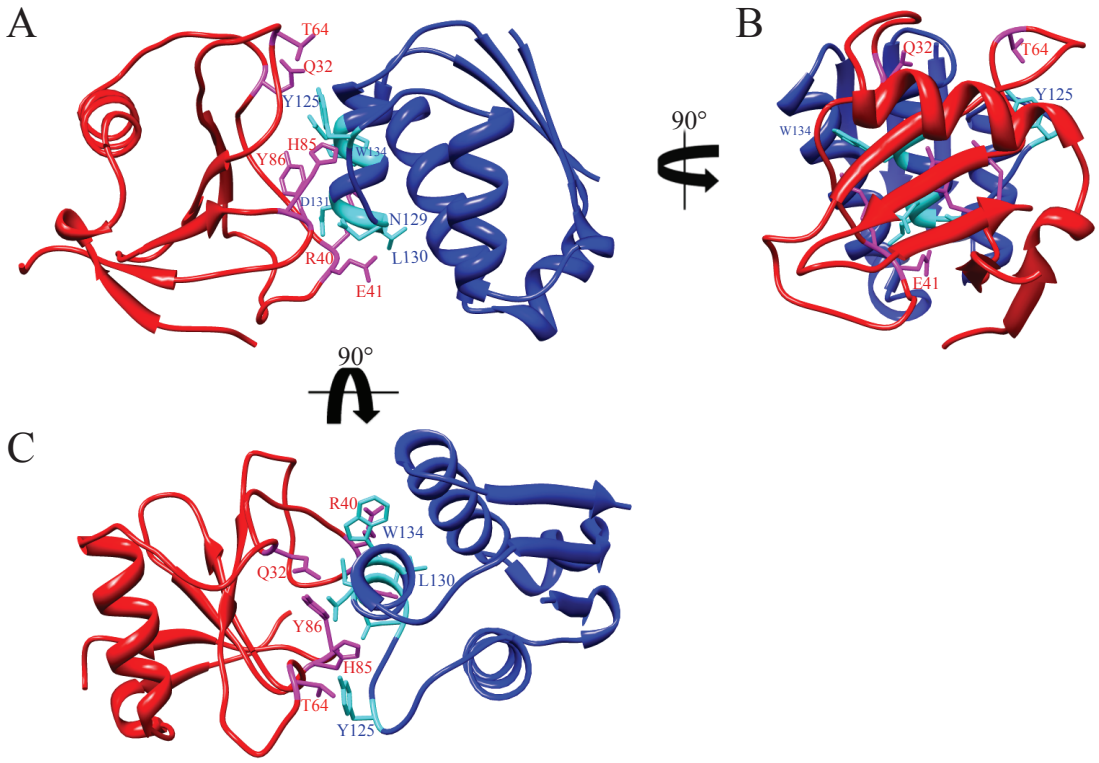

Figure 2.6: The RNase / Barstar complex: Cartoon representation of the wild type RNas (red) in complex with Barstar (blue). RNase and Barstar interface residues depicted in magenta and cyan respectively. A) Front view. B) Side view. C) Top view.

The dissociation constant of the RNase / Barstar complex is higher than Barnase / Barstar by three orders of magnitude due to a reduction of almost 300 $Å^2$ in the buried

surface area of the RNase / Barstar complex. Due to this reduction, the number of hydrogen bonds, salt bridges, trapped waters and the total number of contacts decreased in comparison with the Barnase / Barstar complex.

2.2.1.2.3 The E3 ubiquitin ligase CBL / Ubiquitin complex (2OOB)

E3 Ubiquitin ligase CBL is encoded by Cbl (Casitas B-lineage Lymphoma) mammalian gene. Mutations to this gene have been implicated in a number of human cancers such as acute myeloid leukemia. They are native regulators of receptor tyrosine kinases (RTKs), targeting these enzymes for ubiquitination and degradation. The site where ubiquitin and ubiquitin-like domains (UBL) will bind is the C-terminal ubiquitin-associated (UBA) domain. All UBA domains have a bundle of three α-helices (Figure 2.7 and 2.8) and the majority of UBA domains bind ubiquitin or UBL domains using a hydrophobic path of residues in the α1–α2 loop and helix α3 (Figure 2.7) (Kozlov, Nguyen et al. 2007).

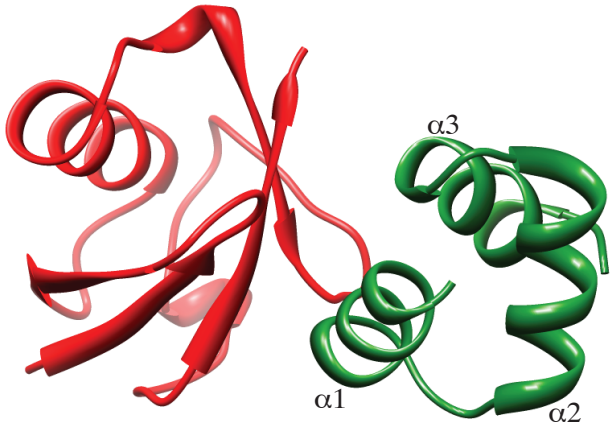

Figure 2. 7: Ubiquitin and Ubiquitin-ligase: Cartoon representation of ubiquitin (red) in complex with E3 ubiquitin ligase Dsk2p (green) (PDB ID: 1WR1).

The UBA domain of Cbl is required for efficient phosphorylation downstream of the EGF and insulin receptor, playing a role in the recruitment of the protein substrate (RTK) (Peschard, Kozlov et al. 2007).

In the 2OOB complex, UBA binds ubiquitin in an unusual way (not observed in other ubiquitin-UBA complexes), using a hydrophobic region in α1 and some residues of α2 (Figure 2.8).

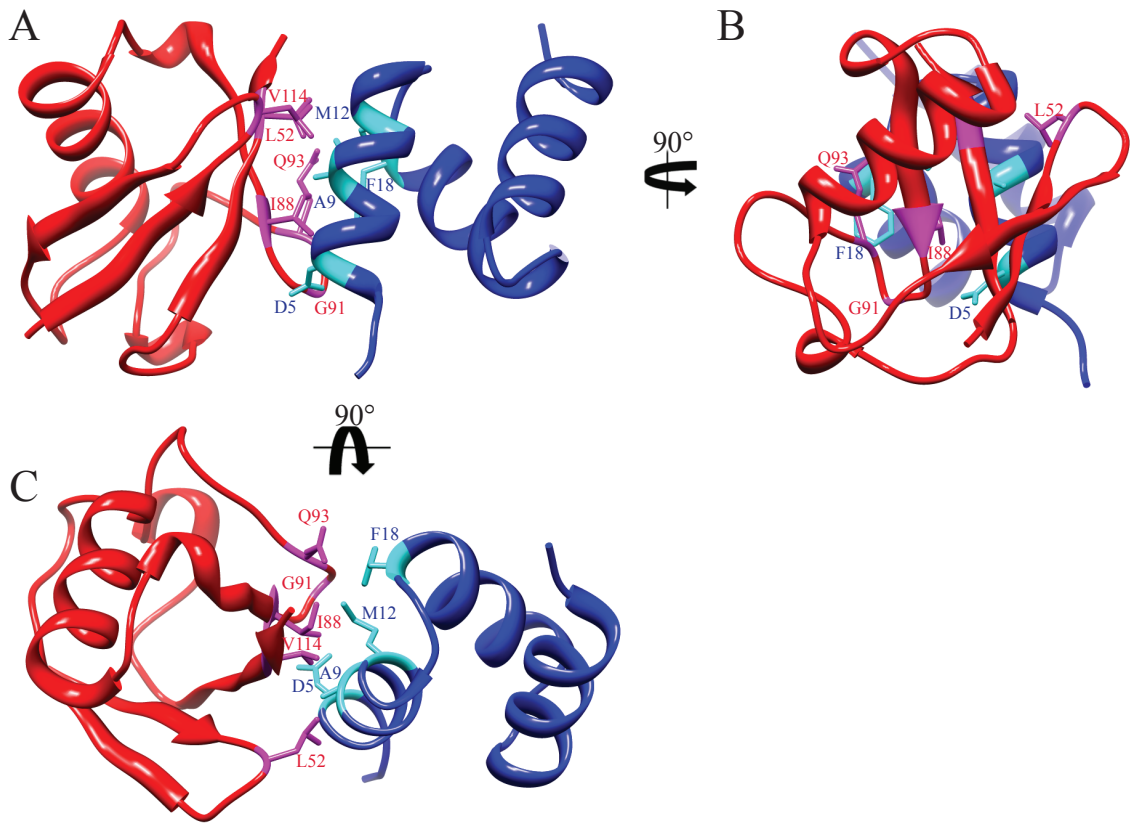

Figure 2.8: The Ubiquitin / Ubiquitin ligase complex: Cartoon representation of Ubiquitin (red) in complex with Ubiquitin ligase (blue). Ubiquitin and Ubiquitin ligase interface residues depicted in magenta and cyan respectively. A) Front view. B) Side view. C) Top view.

2.2.1.2.4 The NucA nuclease – NuiA nuclease inhibitor (2O3B)

NucA is one of the most powerful nucleases known, and it degrades both single and double stranded DNA and RNA. The ability to degrade nucleic acids makes NucA dangerous to the host cell, for that reason the regulation of its activity is critical. NuiA is the natural inhibitor that binds NucA with picomolar affinity. The structure of NuiA is similar to an “open jaw” biting into one side of the NucA molecule (Ghosh, Meiss et al. 2007) (Figure 2.9).

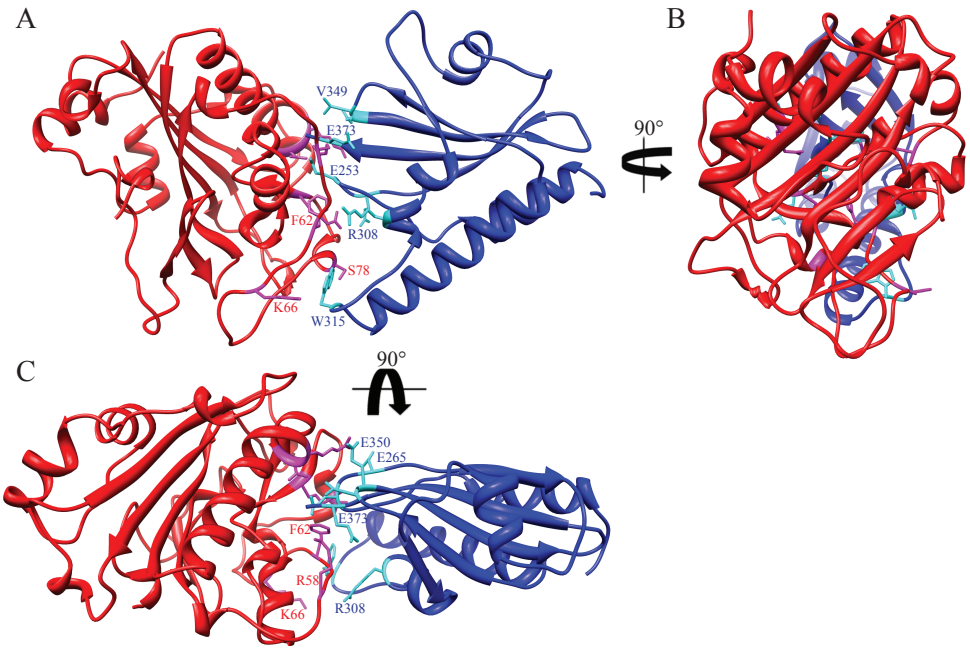

Figure 2.9: The NucA / NuiA complex: Cartoon representation of the wild type NucA (red) in complex with NuiA (blue). NucA and NuiA interface residues depicted in magenta and cyan respectively. A) Front view. B) Side view. C) Top view.

Upon complexation, NucA suffers no significant structural changes in the backbone; it suffers only side chain rearrangements. But NuiA undergoes several backbone conformational changes (Figure 2.10).

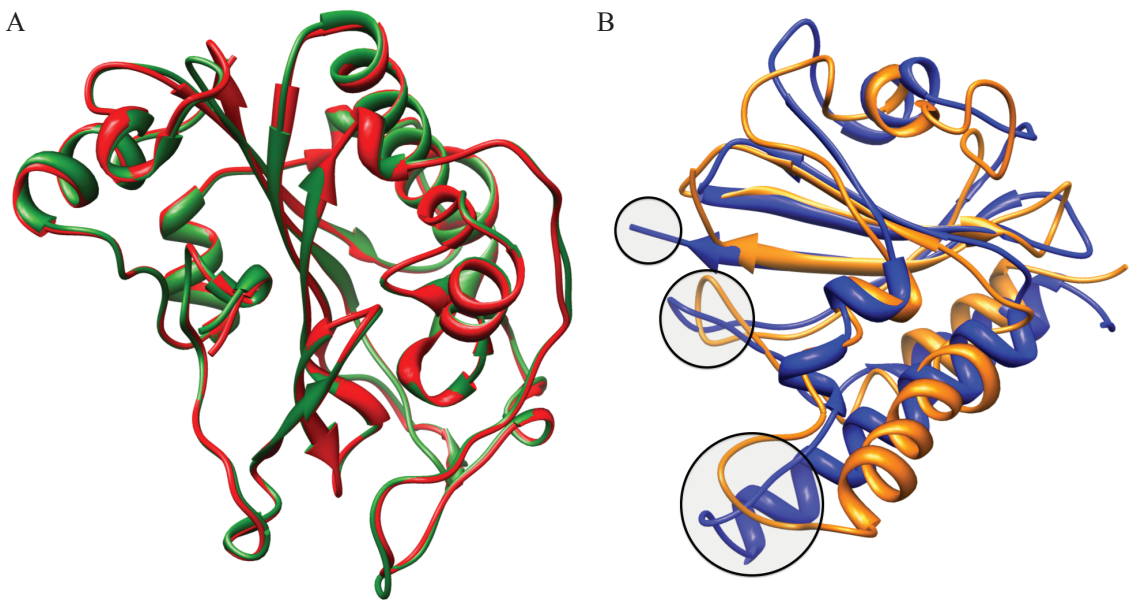

Figure 2.10: Bound versus unbound conformations: Cartoon representation comparing the bound versus the unbound conformations for A) Bound NucA (red) and unbound NucA (green) and B) Bound NuiA (blue) and unbound NuiA (orange). Big conformational changes are encircled.

2.2.2 MD simulations of the ELNEDIN models

Molecular Dynamic (MD) simulations were performed in the NPT ensemble using the GROMACS simulation package version 4.0.7 (Van Der Spoel, Lindahl et al. 2005). A modified Martini 2.1 force field (Marrink, Risselada et al. 2007) was used (see 2.2.2.1). Explicit coarse-grained water molecules and periodic boundary conditions were used to mimic solution conditions. Nonbonded interactions included both Lennard-Jones and Coulomb potentials (cutoff rlist 1.3 nm, rcoul and rvdw 1.2 nm) were computed using the Shift function (Van Der Spoel and Van Maaren 2006). The integration time-step of 20 fs was used and the neighbor list was updated each 5 steps. Berendsen coupling to an external bath (Berendsen, Postma et al. 1984) was used to maintain the pressure (τ_P = 2.0 ps) and temperature constant (τ_T = 0.5 ps) at 1 bar and 300 K respectively.

2.2.2.1 Modified Martini 2.1

A modified Martini 2.1 force field was used. This modified version distinguishes two types of P4 particles: P4 (protein) and A1 (water). The original martini waters (W and WF, water and anti-freezing water, respectively) were P4, BP4, but since other particles were also P4, this affects their interaction with water.

2.2.2.2 Non-bonded parameters for a CG small water

Particles A1 (for W) and BA1 (for WF) from the modified Martini 2.1 were replaced for S1 and BS1, respectively.

The value of sigma (0.32 nm) and epsilon (0.65 $kJmol^{-1}$) for atom OW in the TIP3P and SPC/SPCE water were used as the sigma and epsilon value for the interaction of particle S1 and BS1 with itself.

Geometrical mean combination rules for sigma and epsilon were used to define the cross-term parameters with the other types of bead. The new sigma and epsilon water for CG water is showed in (Table 2.4 and 2.5) respectively.

Table 2.4: Comparison of sigma values: List of the sigma values used with CG regular water and with CG small water.

	Water Particle	Anti-freeze Particle	Other Particles
Regular Water	0.47	0.57	0.47
Small Water	0.32	0.43	0.39

Table 2. 5: Levels of LJ interactions: List of the predetermined values of LJ parameters that indicates the well depth and their respective equivalent for a CG small water.

Level of interaction	ε (prot – A1/BA1)	ε (prot – S1/BS1)
Supra attractive	5.6	1.91
Attractive	5.0	1.81
Almost attractive	4.5	1.71
Semi attractive	4.0	1.61
Intermediate	3.5	1.51
Almost intermediate	3.1	1.42
Semi repulsive	2.7	1.32
Almost repulsive	2.3	1.22
Repulsive	2.0	1.14

2.2.3 Atomistic MD simulations (AT)

The AMBER99SB force field (Hornak, Abel et al. 2006) was used to describe both bonded and non-bonded interactions. Explicit TIP3P waters along with periodic boundary conditions were used to simulate solution conditions. Long range electrostatics were computed using the PME algorithm (Darden, York et al. 1993). Short range non-bonded interactions included both Lennard-Jones and Coulomb potentials (rlist=1.0, rcoulomb 1.0, and rvdw 1.2 nm). The integration time-step of 2 fs was used and the neighbor list was updated each 5 steps. A velocity rescaling thermostat (Bussi, Donadio et al. 2007)

and a Berendsen pressure coupling to an external bath (Berendsen, Postma et al. 1984) were used to maintain the temperature (τ_T = 0.1 ps) and pressure (τ_P = 1.0 ps) constant at 300 K and 1 bar respectively. The LINCS algorithm was used to constrain all bond lengths. The total length of each simulation was 2ns.

2.2.4 Protocol to obtain initial structures for potential of mean force calculations

The general methodology is as follows:

1. Align the principal axis of the complex along the Cartesian coordinates.

2. Only for CG systems, convert each of the proteins in the complex into an ELNEDIN model using a cutoff distance of 1.2 nm and a spring force constant of 1000 $KJ.mol^{-1}.nm^{-2}$. This was done using the MOLSYS library (version 0.3).

3. Translate the ligand from its original position up to 19Å away, along the axis joining the center of mass (COM) of the receptor and the ligand (Figure 2.9). Only the Cα carbons of each protein were used to define the COM. The coordinates of the receptor and the ligand are saved every 0.25Å. The translation was done using MOLSYS (version 0.3).

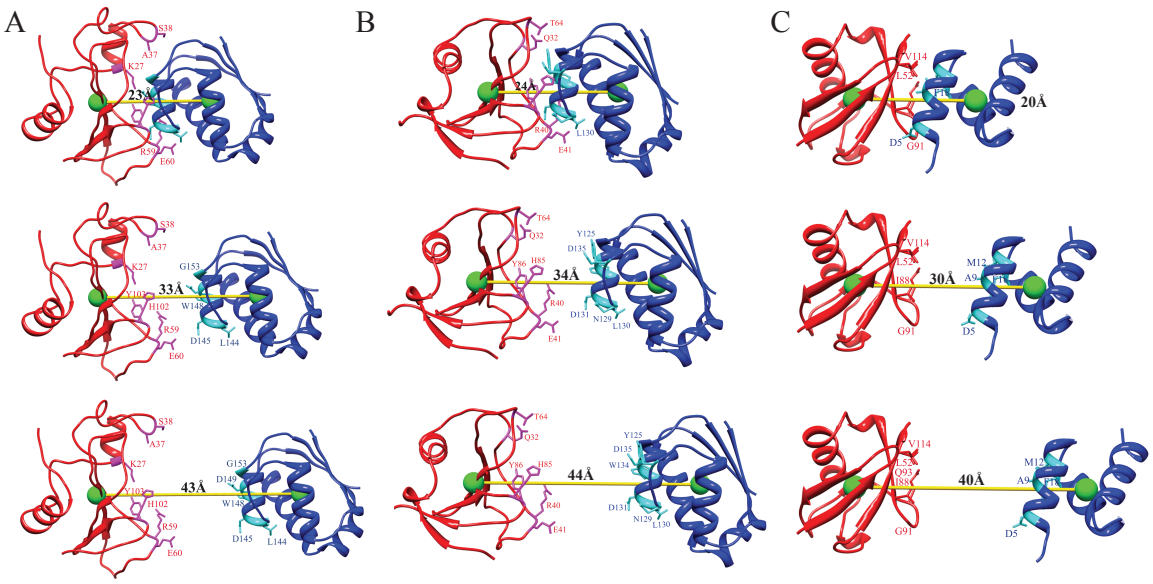

Figure 2.11: Cartoon representation of the translation of the three studied systems: The pictures shows how the ligand was pulled apart from the receptor A) 1BRS, from 23 Å to 43 Å. B) 1AY7, from 24 Å to 44 Å. C) 2OOB, from 20 Å to 40 Å. Even though the performed translation was only 19 Å, the figure shows 20 Å for visualization purposes.

4. Using the same methodology from step 3, bring the ligand closer to the receptor by 1Å, printing structures each 0.25Å. At the end, there should be a total of 79 structures.

5. Place each new structure in a cubic box of water using as a minimum distance between the protein and the box, 1.3 nm, for CG, and 1.0 nm for AT systems. Each water should be placed using the predetermined van der Waals distance of 0.22 nm, for CG waters, and the default value of 0.105 for TIP3P waters.

6. 10% of the total number of CG waters should be CG anti-freezing waters, placed in a systematical way using the MOLSYS 0.3 suite.

7. Perform a steepest descent energy minimization restraining all the backbone beads (BAS) atoms of the protein for 1000 steps for each structure, until the maximum force becomes less than 500 KJ $mol^{-1}nm^{-1}$ (For CG systems) and restraining all heavy atoms of the

protein for 1000 steps, until the maximum force becomes less than 1000 KJ $mol^{-1}nm^{-1}$ (for atomistic systems).

8. Neutralize the charge of each box substituting solvent molecules by monoatomic counterions. This step is done using the tool 'genion' from the GROMACS simulation package version 4.0.7.

9. Repeat step 7.

10. To equilibrate the water and counter-ions molecules, it should be performed, for CG systems, a 250ps -long plain MD simulation restraining all the BAS atoms of the protein, followed by another 250ps removing the position restraints. For atomistic systems, a 5ps -long plain MD simulation restraining all the heavy atoms of the protein.

11. The reaction coordinate is defined by the COM of the BAS atoms of each ligand and receptor, for CG systems. For atomistic systems, the COM is defined by the COM of the Cα atoms of each ligand and receptor.

12. For atomistic systems, the structures obtained, after the 5ps long plain MD simulation restraining all the heavy atoms of the protein, are the initial structures to perform the umbrella simulations to obtain a PMF curve. Each umbrella simulation should be 2ns long using an umbrella force constant of 2000 $kJ.mol^{-1}nm^{-2}$. After umbrella simulations are done, proceed to obtain a PMF curve.

13. For CG systems, perform a series of 250ps -long MD simulations in which the distance between the center of mass of the ligand and receptor is restrained with a harmonic potential, the force constant of which is increased systematically from 1000, 2000, 4000,

8000, 16000 and 32000 kJ.mol^{-1}nm^{-2}. The trajectories of the last set of umbrella simulations should be written each 0.02 ps.

14. For each structure, from the last 50ps of the last set of umbrella simulation, look for the structure with a distance between the COM closest to the reference distance within an error of +/- 0.05 Å. These chosen structures will be the new initial structures to perform umbrella simulations to obtain a PMF curve. If it is not possible to obtain a structure within that error, a new 250ps umbrella simulation must be performed using an umbrella force constant of 64000 kJ.mol^{-1}nm^{-2}.

15. Once a set of 79 new initial structures is obtained, five independent 1250ps umbrella simulations will be performed to each structure changing the initial velocities and using an umbrella force constant of 4000 kJ.mol^{-1}nm^{-2}. These should be repeated four times with different velocities. This way, water-freezing problems will be avoided. By own experience, it was found that after 5ns the water molecules of some systems tend to freeze.

16. After the four sets are ready, for each structure, discard the first 250ps and concatenate the five remaining 1ns. Obtain a PMF curve for each set.

17. Obtain an average PMF (with standard errors) of the four PMF.

2.2.5 Potential of mean force calculations

The corresponding output forces from all the trajectories will be converted to distance using the following relationship:

$$F = -k_{umb} \cdot (x - x_0) \qquad (2.\ 22)$$

Where F=force in KJ/mol^{-1}/nm^{-1}, k_{umb} umbrella constant force and x, x_0 distance and reference distance, respectively.

In order to obtain a PMF curve, it should be performed, to each set of trajectories, the weighted histogram analysis method (WHAM) to reconstruct the full distribution function. This step will be done using the software WHAM 2.0.4 (Grossfield 2003).

3

Free energy profiles of intermolecular recognition based on coarse-grained protein models

3.1 Introduction

Here we seek to test the ability of the ELNEDIN approach to obtain accurate free-energy profiles for the dissociation of binary protein complexes. To this end, three protein-protein complexes were selected based on their affinity: 1) The Barnase / Barstar complex (PDB I.D.: 1BRS), the most common and used complex in literature, this complex presents a high affinity value ($K_d = 2.0x10^{-13}$ / DG = -17.3); 2) The RNase / Barstar complex (PDB I.D.: 1AY7), another Barstar containing complex but of a lower affinity complex ($K_d = 2.0x10^{-10}$ / DG = -13.2) than the previous one; and 3) The Ubiquitin / Ubiquitin Ligase complex (PDB I.D.: 2OOB) a low affinity complex ($K_d = 6.0x10^{-5}$ / DG = -5.7). These three protein-protein complexes were modeled using atomistic and coarse-grained (ELNEDIN) approaches. The ELNEDIN models were built using different network parameters to evaluate the effect of the scaffold's rigidity on the free energy profiles. Rigid scaffolds were built using R_C / k_{SPRING} = 1.2 / 1000 nm / kJ.mol^{-1}.nm^{-2}, moderately flexible scaffolds used a combination of R_C / k_{SPRING} = 1.0 / 500 nm / kJ.mol^{-1}.nm^{-2} and the more flexible scaffolds used a R_C / k_{SPRING} = 0.8 / 500 nm / kJ.mol^{-1}.nm^{-2} combination. Each individual protein model, of different flexibility, was further simulated and the potential of mean force was calculated. The obtained free energy profiles are compared and analyzed.

3.2 Atomistic free-energy profiles

Atomistic approaches (Jorgensen and Tirado-Rives 1988; Cornell, Cieplak et al. 1995; Jorgensen, Maxwell et al. 1996) have shown to be very accurate and able in modeling the protein stability and structure (Gabdoulline and Wade 2001; Spaar and Helms 2005; Spaar, Dammer et al. 2006; Hoefling and Gottschalk 2010; Wang, Siu et al. 2010). The free energy profiles obtained by using atomistic approaches will be used to compare and analyze the free energy profiles obtained later by using the ELNEDIN approach.

Previously, Selzer and Schreiber (Selzer and Schreiber 2001) described two possible scenarios of protein-protein association (Figure 3.1). The two profiles differ in the height of the barriers separating the various systems. In the first case, the rate-determining step is the transition from the encounter complex to the final complex, A:B → AB, (Figure 3.1 A). In the second case the rate determining step is the formation of the encounter complex from the dissociated proteins, A+B → A:B (Figure 3.1 B).

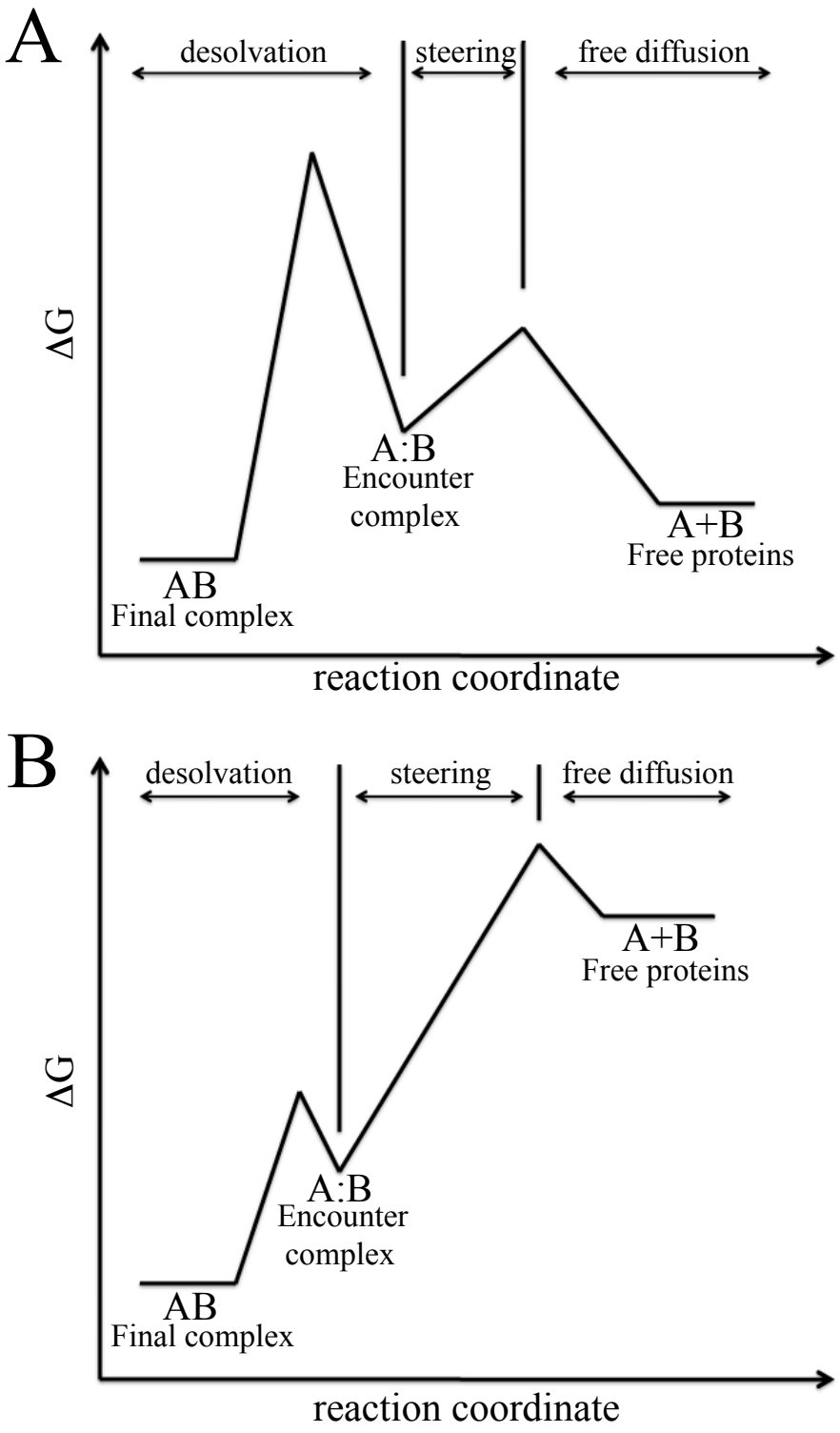

Figure 3.1: Free energy profile describing the protein-protein association pathway: Association pathway suggested by Selzer and Schreiber. A) The rate-determining step is the formation of the final complex. B) The rate-determining step is the formation of the encounter complex.

As reference for all atomistic simulations, standard errors corresponding to twice the standard deviation were calculated from two different free energy profiles coming from two independent sets of MD simulations of the Barnase / Barstar complex. The average standard error is 0.43 ± 0.40 kcal/mol.

Figure 3.2 shows the potential of mean force versus the distance separating the centers of mass of the two proteins for the Barnase / Barstar complex. The minimum of the well correspond to the bound complex AB and at the other extreme A+B represents the dissociated state. There is a non-significant barrier at ~25.5 Å of approximately 0.7 kcal/mol; the height of this barrier falls in the upper limit of the average standard error of 0.43 ± 0.40 kcal/mol. We cannot say that the profile is similar to the second protein-protein association profile proposed by Seltzer and Schreiber (Selzer and Schreiber 2001) (Figure 3.1 B), because the rate-limiting barrier between the dissociated state, A+B, and the encounter complex, A:B, is not observed in our results, and the desolvation barrier is not significant. The computed free energy of binding is -18.3 kcal/mol while the experimental value is -17.3 kcal/mol.

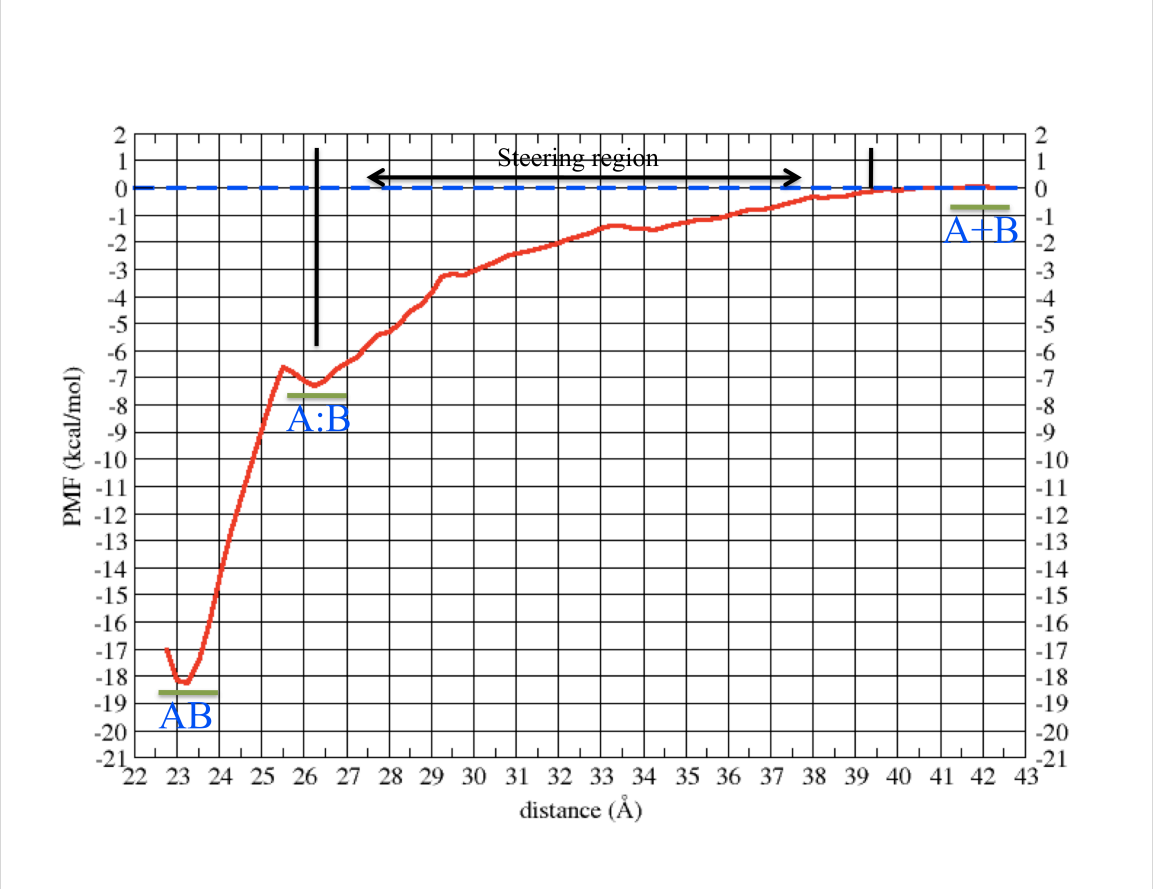

Figure 3.2: Free energy profile of the association of Barnase / Barstar.

Other authors have obtained profiles similar to ours: Hoefling and Gottschalk (Hoefling and Gottschalk 2010) using the OPLS force field, Wang et al (Wang, Siu et al. 2010) using the GROMOS 96 53a6 force field, Neumann and Gottschalk (Neumann and Gottschalk 2009) using steered molecular dynamics, Spaar et al (Spaar and Helms 2005; Spaar, Dammer et al. 2006) using Brownian dynamics in combination with the AMBER95 force field. All of these authors have obtained the minimum of the well very close to what we obtained. These groups also saw the location of a non-significant barrier before the formation of the final complex, and it is believed to be the displacement of bound water from the interface (Janin 1997; Camacho, Weng et al. 1999; Selzer and Schreiber 2001; Hoefling and Gottschalk 2010). The size of this barrier

depends on the concentration of ions in the system and ranges between 0.3 to 2.0 Å (Selzer and Schreiber 2001; Spaar, Dammer et al. 2006; Hoefling and Gottschalk 2010; Wang, Siu et al. 2010); because this barrier is not significant, it is not the rate-limiting step, corroborating the classical view that the transition from the encounter complex to the final complex is fast and not rate limiting (Janin 1997; Camacho, Weng et al. 1999; Gabdoulline and Wade 1999). For that reason, for some authors (Janin 1997; Hoefling and Gottschalk 2010; Wang, Siu et al. 2010), the overall process is considered a downhill one where the complex goes from a A+B state to a AB state without any significant barrier. This barrier depends on the electrostatics and was observed by Hoefling, when an ionic concentration of 150mM is used, and by Wang for the Barnase / Barstar complex, when an ionic concentration of 100mM is used. What differs among all these authors is the free energy of binding: For Hoefling et al is -19.1 kcal/mol when only counter-ions are present and -11.2 kcal/mol in 150 mM of ionic strength. For Wang is -12.9 kcal/mol using an ionic strength of 100 mM. Neumann et al (Neumann and Gottschalk 2009) have an error of a factor of 2.5 – 3 kcal/mol because Barstar tends to frequently unfold during the steered MD simulation. Spaar et al present a free energy binding of -4 kcal/mol (Spaar and Helms 2005; Spaar, Dammer et al. 2006). In all of these cases it is possible to see that the choice of the force field, or the technique used to obtain the free energy profile, plays a role in the computation of the absolute free energy values, but that the general shape is not strongly influenced.

The profile of the RnaseA / Barstar complex (Figure 3.3) is also very similar to the Barnase / Barstar one (Figure 3.2); the minimum of the bound state (AB) is at ~24.75 Å with a non-significant barrier at 27 Å of ~0.6 kcal/mol and a free energy of binding of -9.7 kcal/mol when the experimental value is -13.2 kcal/mol (Figure 3.2 B). In the Barnase / Barstar and RNase / Barstar complexes, it is possible to observe the large size of the steering region (Figure

3.2 and 3.3), suggesting that strong electrostatic forces are steering these complexes (Camacho, Weng et al. 1999; Selzer and Schreiber 2001), a fact that is already known by the structure of Barnase / Barstar (Buckle, Schreiber et al. 1994) and RnaseA / Barstar (Sevcik, Urbanikova et al. 1998).

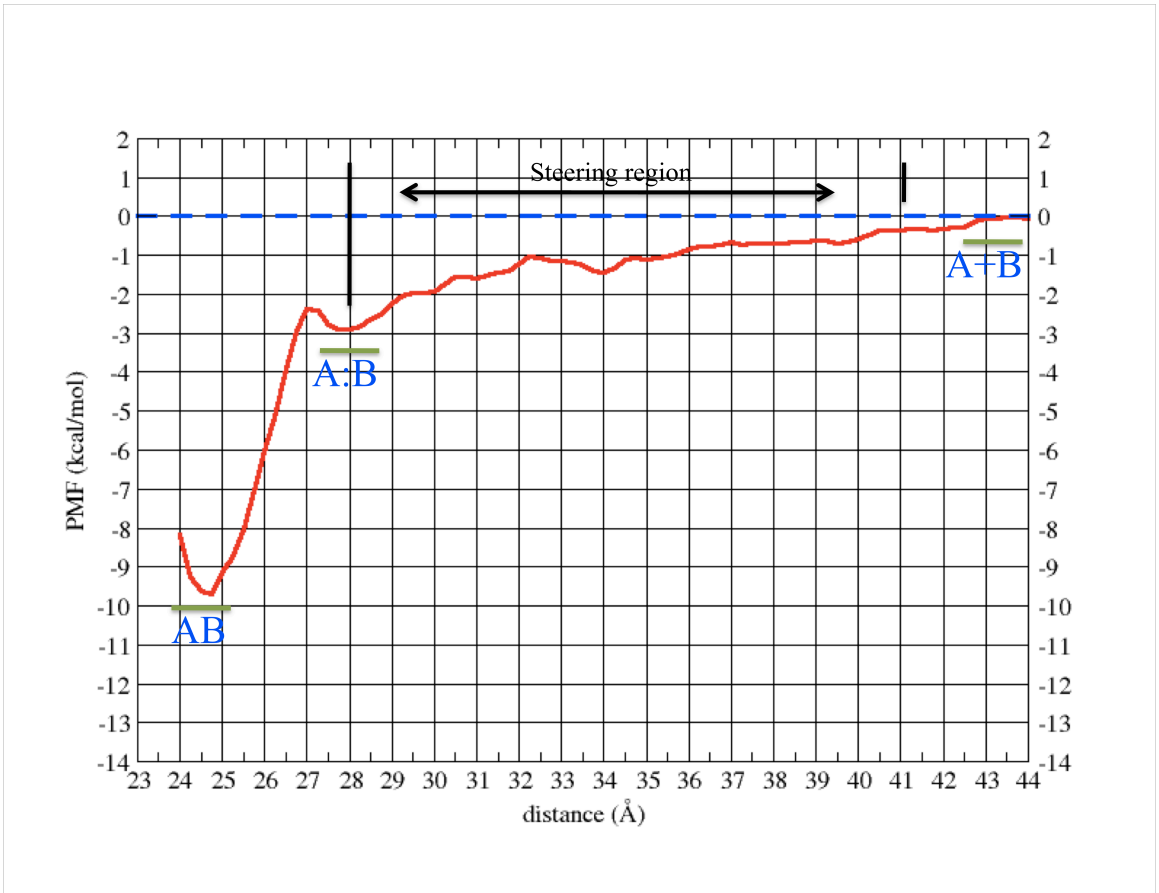

Figure 3.3: Free energy profile of the association of RNaseA / Barstar.

The ubiquitin – ubiquitin ligase free energy profile (Figure 3.4) shows two possible barriers of less than 0.5 kcal/mol, this value is not significant because it falls inside the average standard error of 0.43 ± 0.40 kcal/mol and it is lower than the order of one k_BT (0.60 kcal/mol). The small size of this non-significant barrier after the formation of the encounter complex might

suggest the absence of desolvation because the interface is very hydrophobic, confirming what is already known: that this complex is attracted by hydrophobic forces (Kozlov, Nguyen et al. 2007; Kozlov, Peschard et al. 2007; Peschard, Kozlov et al. 2007). Selzer (Selzer and Schreiber 2001) also suggest that a very small steering region might indicate that the electrostatic forces are weak. The free energy of binding is -9.0 kcal/mol when the experimental value is -5.7 kcal/mol.

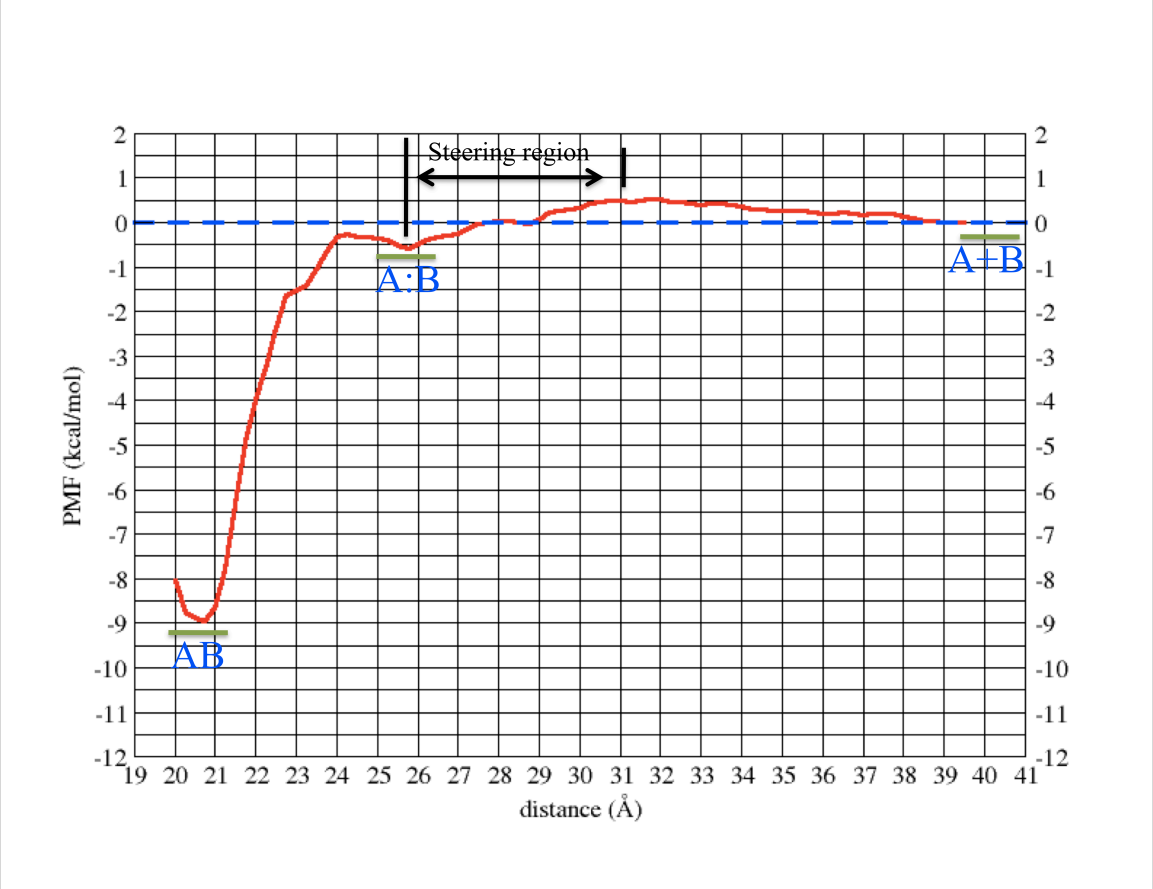

Figure 3.4: Free energy profile of the association of Ubiquitin / Ubiquitin ligase.

3.3 Coarse-grained free-energy profiles

3.3.1 Rigid or Flexible networks?

Previously, our laboratory has shown that a flexible scaffold of R_C / k_{SPRING} = 0.8 / 500 nm / kJ.mol^{-1}.nm^{-2} is good for reproducing dynamical properties, but the preliminary studies on protein-protein association showed that the choice of the network could play a determining role in complex formation. Many authors suggest that in protein folding the protein should be flexible to allow the conformational changes, but in protein-protein association, the two proteins have to have a rigid structure to facilitate the initial encounter and the mutual alignment (Janin 1997; Selzer and Schreiber 2001; Chen and Weng 2002; Chen, Li et al. 2003; Chen, Mintseris et al. 2003).

We have used three networks of varied flexibility: a) rigid (R_C = 1.2 nm and K_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} or 1.2/1000); b) intermediate (R_C = 1.0 nm and K_{SPRING} = 500 kJ.mol^{-1}.nm^{-2} or 1.0/500); and c) flexible (R_C = 0.8 nm and K_{SPRING} = 500 kJ.mol^{-1}.nm^{-2} or 0.8/500). Standard errors are twice the standard deviation that was calculated from the 4x5ns MD simulations.

Figure 3.5 A corresponds to the potential of mean force versus the distance of separation between the center of masses of the two proteins for the Barnase / Barstar complex using the three different elastic networks. The profile of the most rigid network (green) looks similar to the one obtained with the atomistic approach (Figure 3.5 B). In Figure 3.5 A, the barrier between the encounter formation and the final complex exists only for the most rigid network (green), this barrier is significant, and it is greater than

the average standard error of 0.25 ± 0.22. Network 1.2/1000 (green) presents a very well defined minimum (AB). The other two networks look very similar, the only difference is in the region of the bound complex 1.0/500 (orange) is a broad minimum while 0.8/500 (maroon) the minimum is not well defined; this data suggests that the R_C parameter might control the shape of the well, but the shape of the profile is strongly influenced by the K_{SPRING} value. Figure 3.5 B is showing the comparison of the free energy profile obtained for the Barnase / Barstar complex by using the elastic network of 1.2/1000 with the ones obtained by using the atomistic approach is shown in the Figure 3.8. The free energy profile of both of them looks very similar. The absolute minimum of the free energy profile obtained with the rigid elastic network is shifted to the right in the reaction coordinate axis by less than 1.5 Å due to deviations from the AMBER99SB approach after minimization and water equilibration. These deviations depend on the structural mapping of the beads in the CG approach that are bigger than the size of atoms in the AMBER99SB approach.

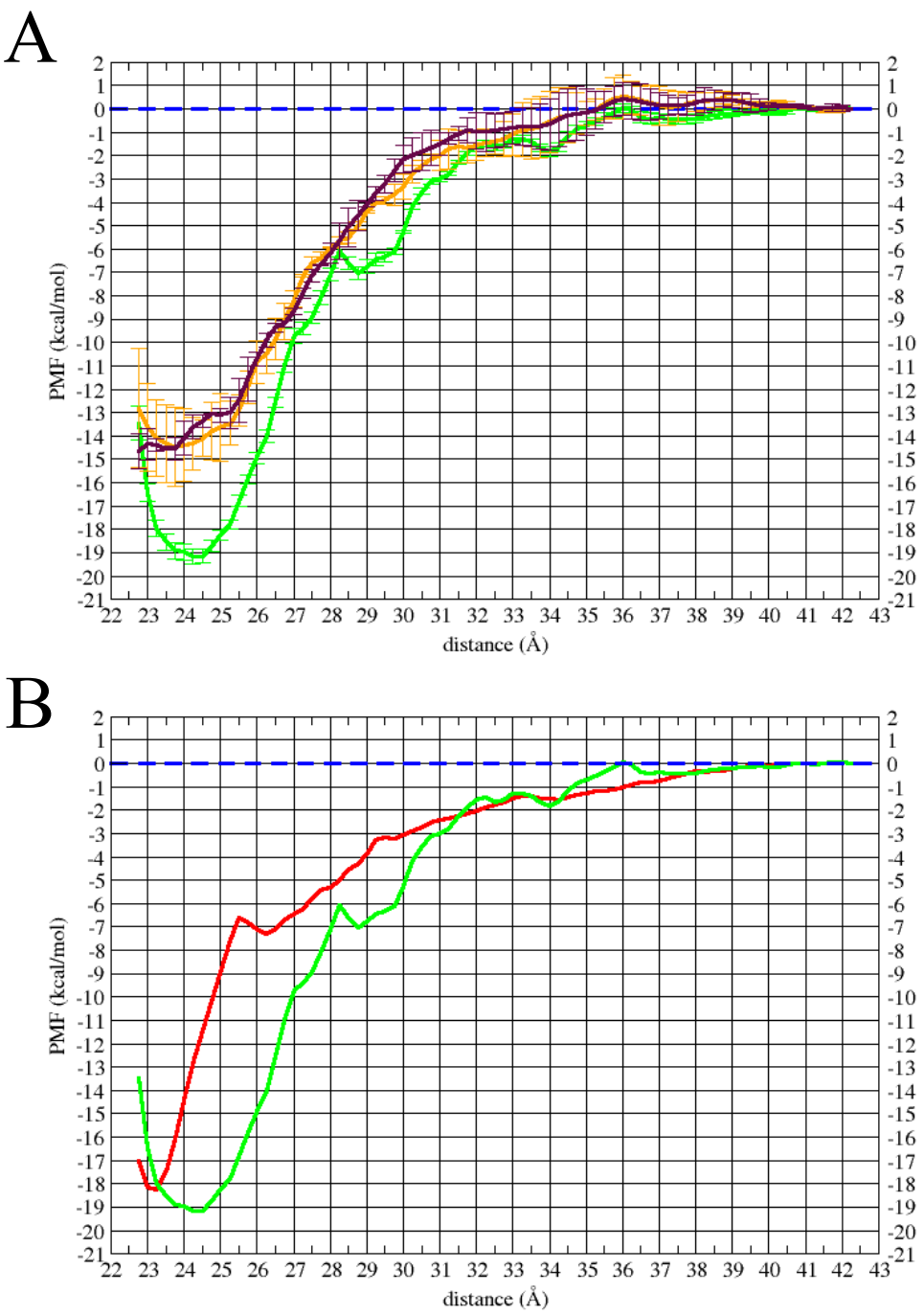

Figure 3.5: The Potential of mean force of the dissociation of the Barnase / Barstar complex: A) Obtained by different ELNEDIN scaffolds: Network 1.2/1000 is depicted in green, 1.0/500 in orange, and 0.8/500 in maroon. B) Free energy profile comparison between the atomistic (red) and the ELNEDIN (1.2/1000) (green) approaches.

The free energy profiles of RNaseA / Barnase obtained using the three elastic networks are shown in figure 3.6 A. The profile of the most rigid network looks similar to the one obtained with the atomistic approach (Figure 3.6 B). In the profile obtained using the rigid elastic network of 1.2/1000, the non-significant barrier after the formation of the encounter complex fell inside the range of our average standard error of 0.35 ± 0.25. The free energy profile is also shifted in the reaction coordinate axis by ~1 Å. Spaar et al (Spaar, Dammer et al. 2006) also obtained a very small barrier for the Barnase / Barstar complex, similar to the one obtained for the rigid elastic network; they termed this a 'shoulder', a more detailed study suggest that this 'shoulder' is indeed a barrier that divides the favorable energetic region near the RNA binding loop and the region of the encounter complex.

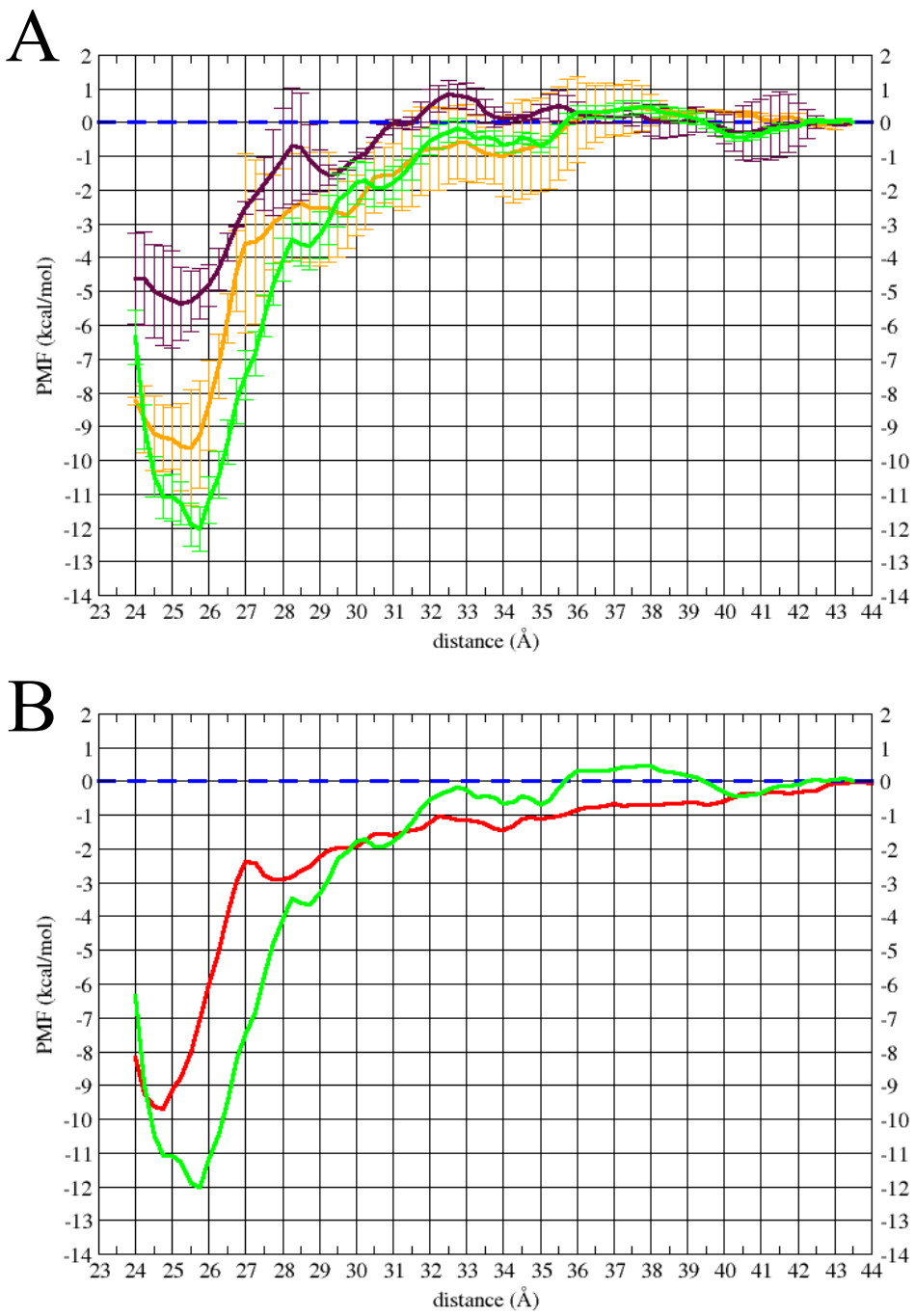

Figure 3.6: The Potential of mean force of the dissociation of the RNaseA / Barstar complex: A) Obtained by different ELNEDIN scaffolds: Network 1.2/1000 is depicted in green, 1.0/500 in orange, and 0.8/500 in maroon. B) Free energy profile comparison between the atomistic (red) and the ELNEDIN (1.2/1000) (green) approaches.

The free energy profiles of Ubiquitin / Ubiquitin ligase obtained using the three different elastic networks are shown in figure 3.7 A. The free energy profile obtained with the elastic network of 1.2/1000 looks similar to the one obtained with the atomistic approach. The barrier after the formation of the encounter complex is present only in the most rigid (green) network, this barrier (~0.6 kcal/mol) is non-significant because it is inside the range of the average standard error of 0.37 ± 0.27. Networks 1.0/500 (orange) and 0.8/500 (maroon) present a broad barrier in the steering region of ~1 kcal/mol. It is not significant because it is within the average standard error of 0.44 ± 0.72 for 1.0/500 and 0.62 ± 0.56 for 0.8/500. The PMF profile is also shifted in the reaction coordinate axis by ~1 Å.

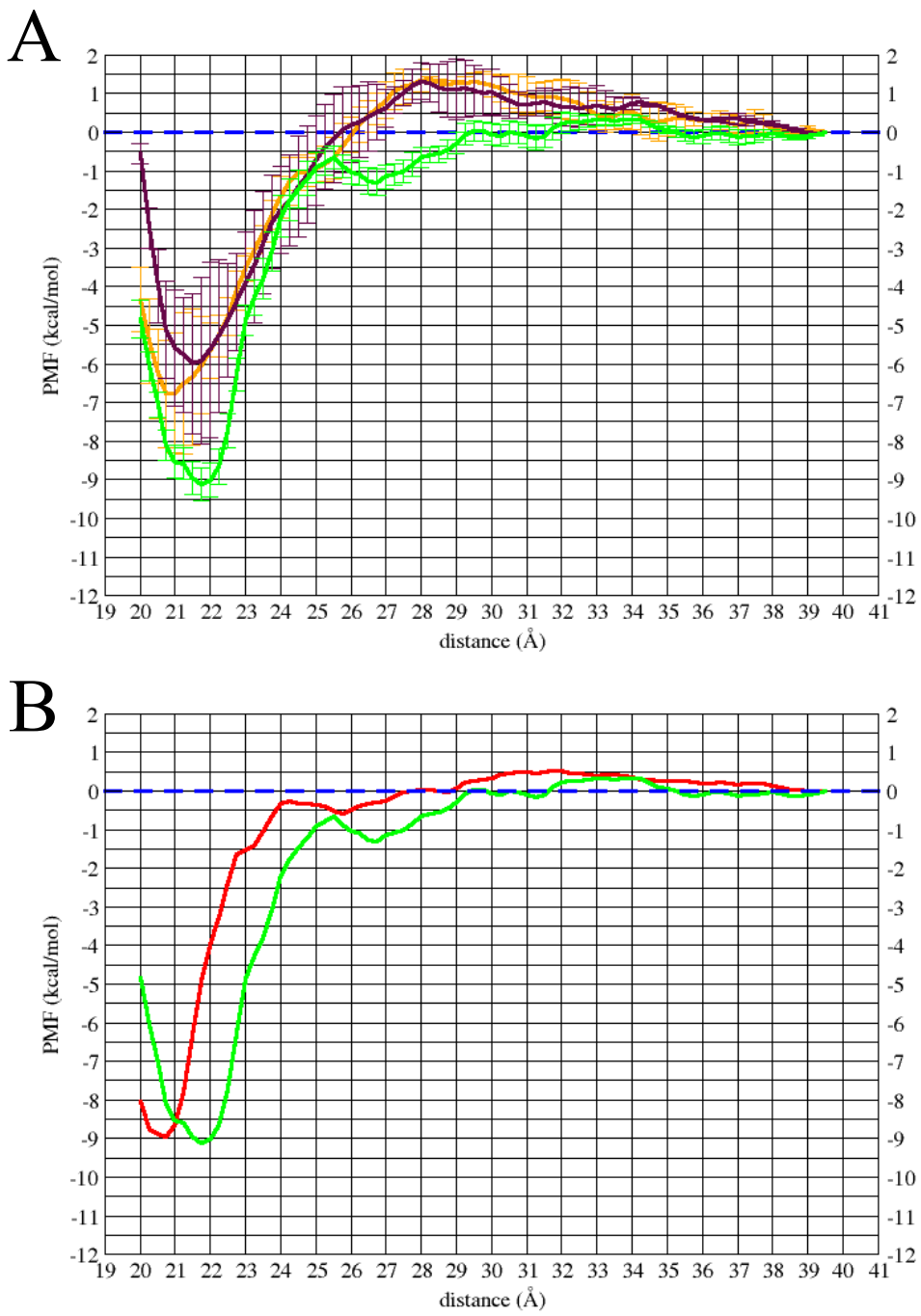

Figure 3.7: The Potential of mean force of the dissociation of the Ubiquitin / Ubiquitin ligase complex: A) Obtained by different ELNEDIN scaffolds: Network 1.2/1000 is depicted in green, 1.0/500 in orange, and 0.8/500 in maroon. B) Free energy profile comparison between the atomistic (red) and the ELNEDIN (1.2/1000) (green) approaches.

The results presented in Figure 3.5 - 3.7 show that the free energy profile obtained using the most rigid elastic network (green) presents the same characteristics as the atomistic approach. In the three cases, the general free energy profile shape is not strongly influenced by the choice of the force field.

The comparison between the computed free energy value obtained using the elastic network 1.2/1000 and the experimental free energy values (Table 3.1) shows that for 1BRS and 1AY7 there is a difference less than 2 kcal/mol, but the difference for 2OOB is almost 4 kcal/mol. It cannot be a problem of the ELNEDIN approach because the result obtained also by the atomistic approach is very close to the value obtained by using the 1.2/1000 scaffold.

Table 3.1: ΔG values comparison: List of the ΔG values obtained experimentally and calculated using the AMBER99SB and the CG/ELNEDIN approach.

Complex	ΔG_{exp} (kcal/mol)	$\Delta G_{AMBER99SB}$ (kcal/mol)	$\Delta G_{1.2/1000}$ (kcal/mol)	$\Delta G_{1.0/500}$ (kcal/mol)	$\Delta G_{0.8/500}$ (kcal/mol)
1BRS	-17.3	-18.3	-19.2	-14.5	-14.7
1AY7	-13.2	-9.7	-12.0	-9.6	-5.4
2OOB	-5.7	-9.0	-9.1	-6.8	-6.0

These results suggest that the rigid network of 1.2/1000 is the best, because it conserves the details of the atomistic profile but shifts the location of the minimum to the right.

3.3.2 ELNEDIN shows better profile than using atomistic models with a low quality electrostatic treatment

We have compared the free energy profiles obtained with ELNEDIN 1.2/1000, against free energy profiles obtained with the atomistic approach using a different electrostatic treatment: a cutoff of 0.8 nm, a PME interpolation order of 4 and a Fourier grid-spacing of 0.12 nm or 0.8/4/0.12. Our default long-range electrostatic treatment for the atomistic approach uses a cutoff of 1.0 nm, a PME interpolation order of 6 and a grid-spacing of 0.10 nm or 1.0/6/0.10. Figure 3.18 is showing the comparison of the Barnase / Barstar free energy profiles obtained with the atomistic approaches using the long-range electrostatic treatment of 1.0/6/0.10 (red), 0.8/4/0.12 (black) and the elastic network of 1.2/1000 (green). The results show that the desolvation barrier between the encounter and the final complex is smaller and not well defined in the atomistic free energy profile obtained with the 0.8/4/0.12 electrostatic treatment. The difference of the free energy of binding between the atomistic approach using an electrostatics of 0.8/4/0.12 and the one obtained with the 1.2/1000 elastic network is greater than the one obtained with the atomistic approach using a higher electrostatic treatment.

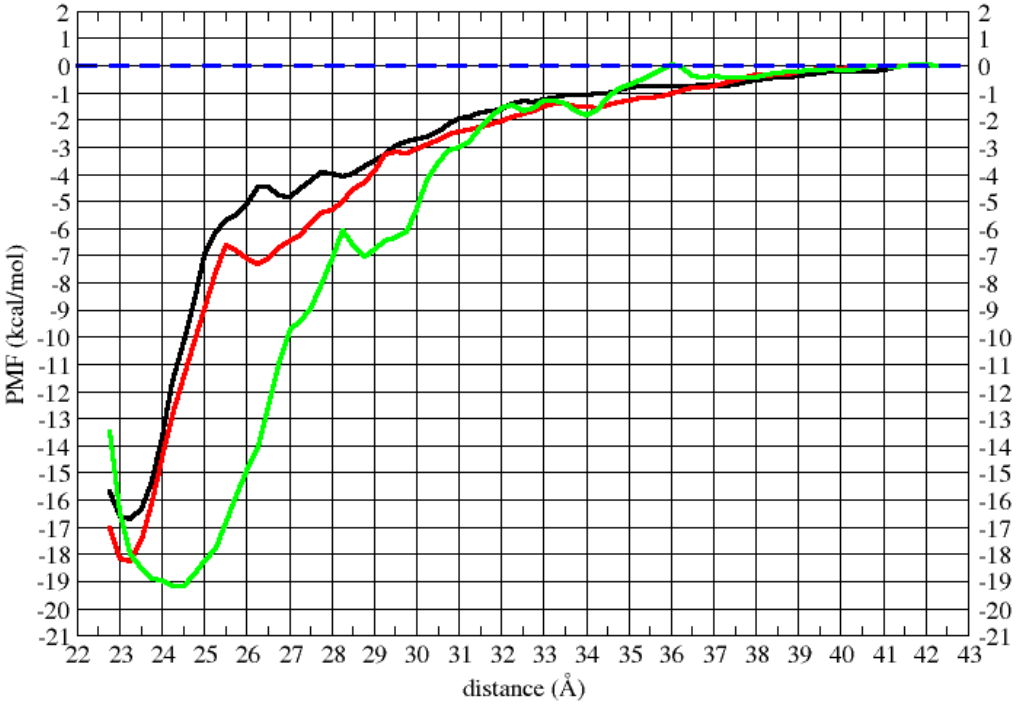

Figure 3.8: Comparison of the Potential of mean force of the dissociation of the Barnase / Barstar complex: PME 0.8/4/0.12 is in black, PME 1.0/6/0.10 is in red, and ELNEDIN 1.2/1000 is depicted in green.

Figure 3.9 details the comparison of the RNaseA / Barstar free energy profiles obtained by using the atomistic approaches with an electrostatic treatment of 1.0/6/0.10 (red), 0.8/4/0.12 (black) and the ELNEDIN approach with an elastic network of 1.2/1000 (green). The results show that the barrier between the encounter and the final complex is smaller, broader and not well defined in the low quality electrostatic treatment. The difference of the free energy of binding between the atomistic approach using a electrostatic of 0.8/4/0.12 and the one obtained with the 1.2/1000 elastic network is

greater than the one obtained with the atomistic approach using a higher electrostatic treatment.

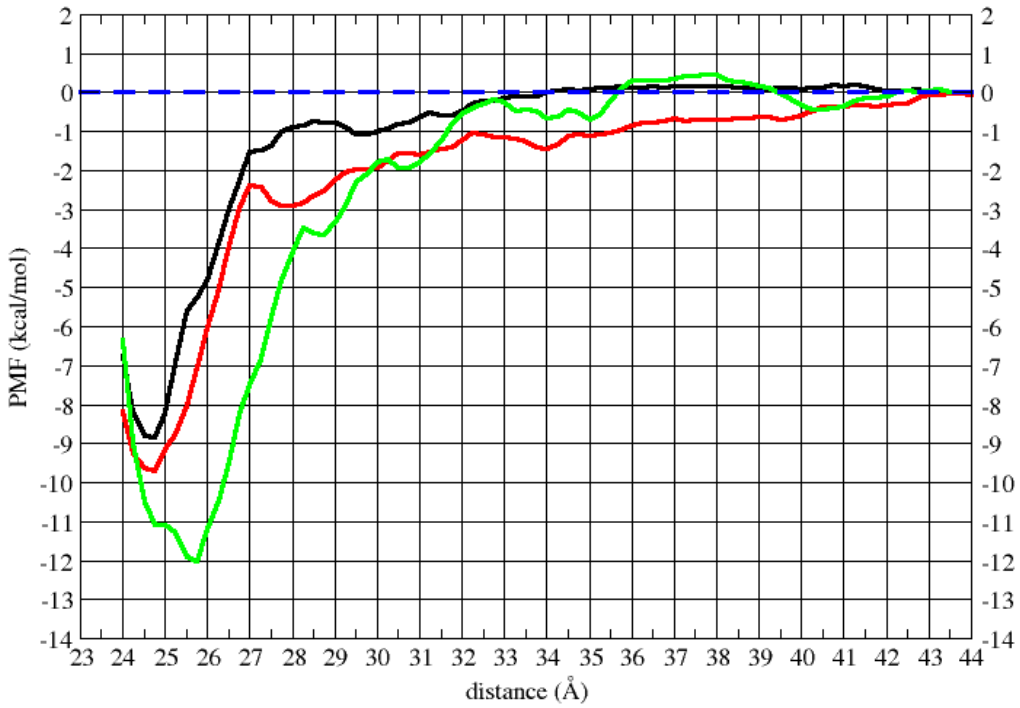

Figure 3.9: Comparison of the Potential of mean force of the dissociation of the RNaseA / Barnase complex: PME 0.8/4/0.12 is in black, PME 1.0/6/0.10 is in red, and ELNEDIN 1.2/1000 is depicted in green.

Figure 3.10 is showing the comparison of the Ubiquitin / Ubiquitin ligase free energy profiles obtained by using the atomistic approaches with an electrostatic treatment of 1.0/6/0.10 (red), 0.8/4/0.12 (black) and the ELNEDIN approach with an elastic network of 1.2/1000 (green). The results show that the barrier between the encounter and the final complex is not present in the low quality electrostatic treatment. The difference

of the free energy of binding between the AMBER99SB 0.8/4/0.12 and the CG is greater than the AMBER99SB 1.0/6/0.10.

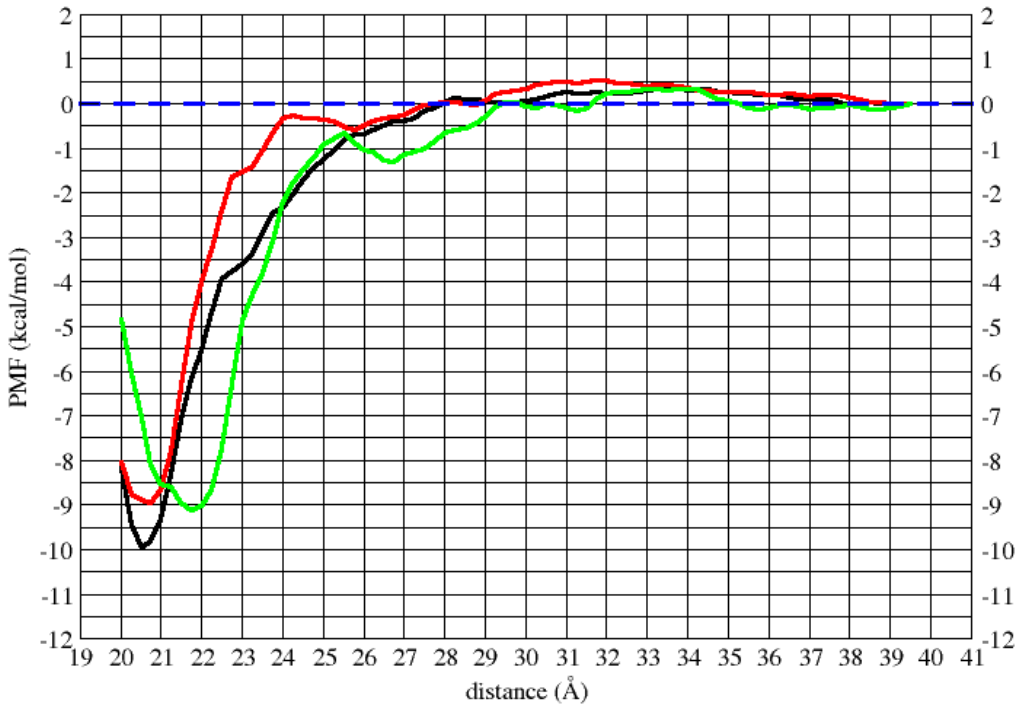

Figure 3.10: Comparison of the Potential of mean force of the dissociation of the Ubiquitin / Ubiquitin ligase complex: PME 0.8/4/0.12 is in black, PME 1.0/6/0.10 is in red, and ELNEDIN 1.2/1000 is depicted in green.

These results further support the idea of how well the ELNEDIN approach maintains the quality of the free energy profile.

4

The protein recognition energy landscape

4.1 Introduction

The first objective of this chapter is to see if the free energy surface obtained when the receptor and the ligand are initially aligned in the correct position has unique characteristics vs. when the ligand is not initially oriented properly or is facing the receptor with a different surface other than the binding surface. For this goal we will use the Barnase / Barstar complex. The ligand will be rotated around the receptor. For each generated configuration, we will perform potential of mean force calculations which will be used to reconstruct the free energy surface.

The second objective is to test the effect of the solvent on the shape of the free energy profile. In order to achieve this goal we will use the same three protein complexes used in chapter 3: 1) The Barnase / Barstar complex (PDB I.D.: 1BRS), 2) The RNase / Barstar complex (PDB I.D.: 1AY7), and 3) The Ubiquitin / Ubiquitin Ligase complex (PDB I.D.: 2OOB). The effect of the solvent will be tested in two ways: 1) by modifying the level of the Lennard-Jones potentials for the interaction of protein – solvent by making it more attractive or repulsive. 2) By replacing the coarse-grained water by smaller coarse-grained water.

4.2 The shape of the free energy surface around a protein receptor

We have chosen as a model system the Barnase / Barstar complex. The initial structure of the Barnase / Barstar complex is the X-ray crystal structure where the principal axis of the complex was aligned along the x-axis. The original position of the ligand (Figure 4.1, blue) was rotated in intervals of 15° around the z-axis, having as rotation center, the center of mass of the receptor (Figure 4.1, red). Each generated structure was checked for possible clashes, the clashes were then removed by translating the ligand away from the receptor along the reaction coordinate 0.1 Å at a time until the average overlap was less than 0.22 Å / atom and no more than 30 pairs of atoms formed clashes; these cutoff values were obtained from averaging the overlap values per atom and the total number of clashes for the 43 complexes obtained from the Protein Docking Database Benchmark 4.0 (see methods 2.2.1.1). Once all the structures were generated, we proceeded to perform umbrella molecular dynamics simulations and potential of mean force calculations following the protocol described in sections 2.2.4 and 2.2.5.

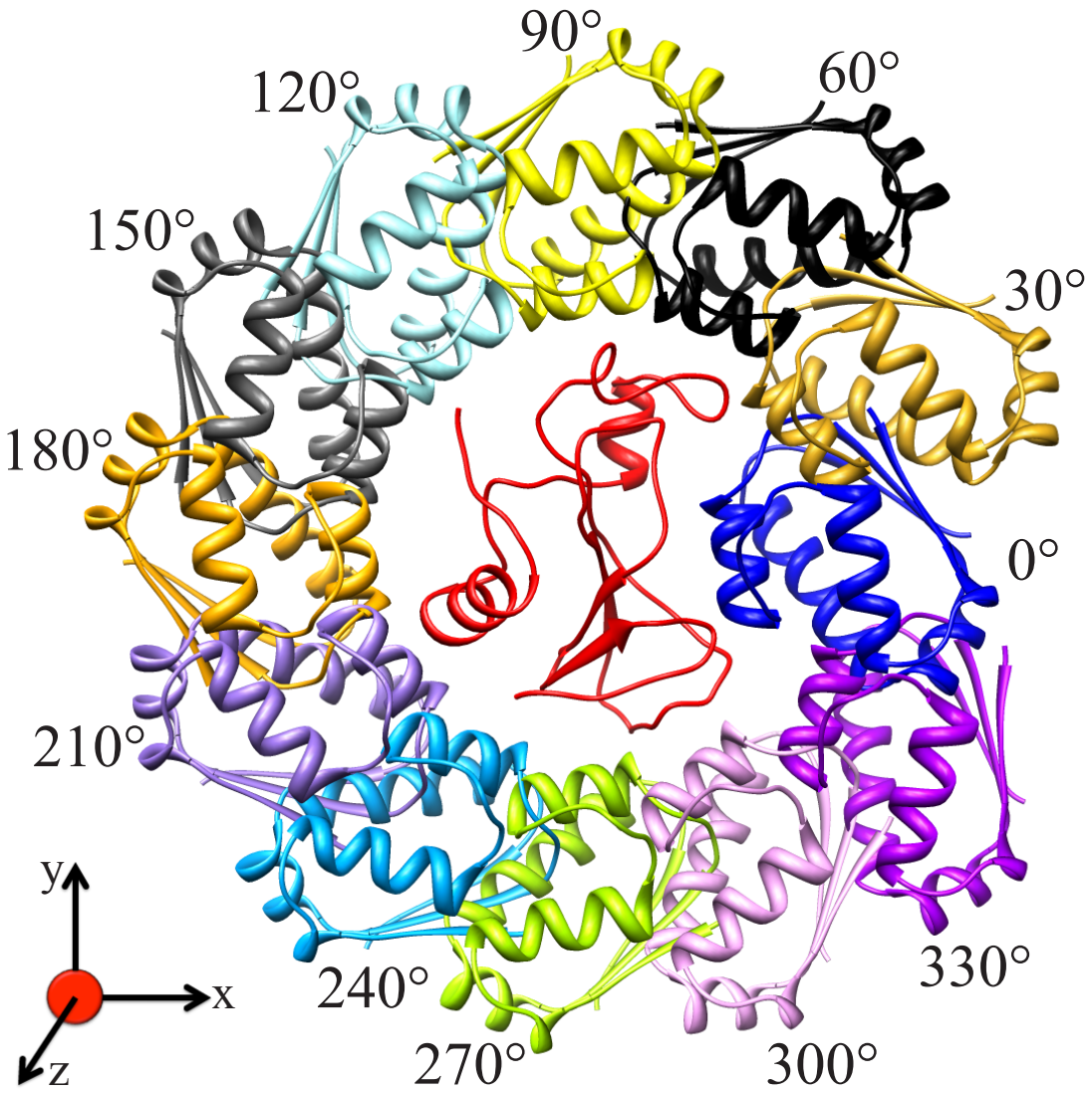

Figure 4.1: Barnase / Barstar conformational search: Rotations performed each 15° around the center of mass of the receptor (Barnase in red). For visualization purposes the ligand structures are plotted each 30° only.

The result (Figure 4.2) shows that the free energy surface has the shape of a funnel, where the bottom of the funnel is the global and lowest minimum and it starts to increase smoothly. Zhang et al showed the presence of energy gradients or funnels near the binding sites (Camacho, Weng et al. 1999; Zhang, Chen et al. 1999). The free energy landscape presents only

one deepest local minimum with a free energy value of ~-19 kcal/mol. The presence of this deepest minimum is in agreement with some authors (Tsai, Kumar et al. 1999; Zhang, Chen et al. 1999) who suggest that the conformational diversity of rigid complexes is limited, and its minimum is single.

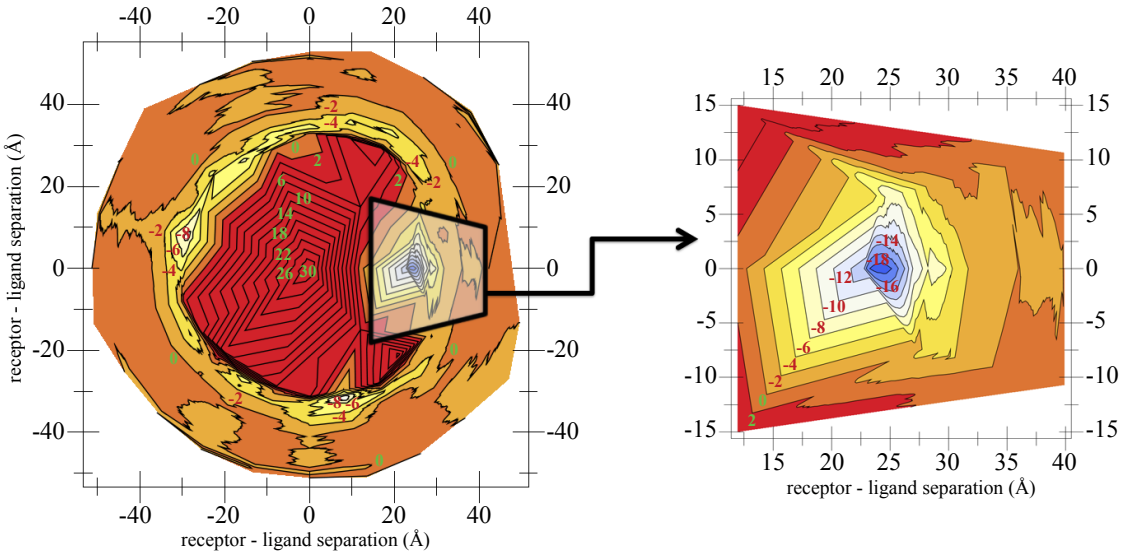

Figure 4.2: Free energy map for the Barnase / Barstar complex: Each contour layer on the free energy surface represents 2 kcal/mol.

The ligand was rotated about its own center of mass each 60° around the x and y axes (Figure 4.3). When the ligand is rotated around the x-axis, the binding surface is always facing the receptor. But when it is rotated around the y-axis, the receptor will face surfaces other than the binding surface.

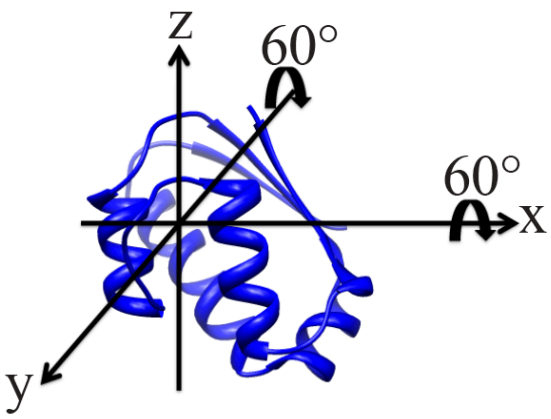

Figure 4.3: Rotating the ligand on its own axis: Additional 60° rotations performed to native Barstar around the x- and y-axis on its own center.

The same procedure, explained in section 4.2, was followed. Each new initial structure, obtained by rotating the X-ray crystal structure position of the ligand, was further rotated along the z-axis, having as a rotation pivot the center of mass of the receptor. Umbrella molecular dynamics simulations and potential of mean force calculations were obtained following the steps described on section 2.2.4 and 2.2.5.

Figure 4.4 shows that the free energy surface obtained when the initial structure is at 0° (Figure 4.4 A) has a characteristic free energy landscape different than the ones obtained when the initial structure at position 0° where the ligand was rotated on its own axis around x (Figures 4.4 B - F). The next free energy landscape with a very close shape to the one showed in Figure 4.4 A corresponds to the free energy surface when the initial structure is rotated 60° and 300° (-60°) (Figure 4.4 B and F, respectively) but with a less deep global minimum. This result suggests that the ligand can rotate from 60° to -60°, and still bind to the receptor.

The free energy profiles obtained when the initial structure was rotated 240° (-120°) (Figure 4.4 E) show that there is still some trace of the energy gradient or funnel. The free energy

profiles obtained when the initial structure was rotated 120° and 180° (Figure 4.4 C and D, respectively) only shows a bunch of crevices and bumps with local minima along the free energy map.

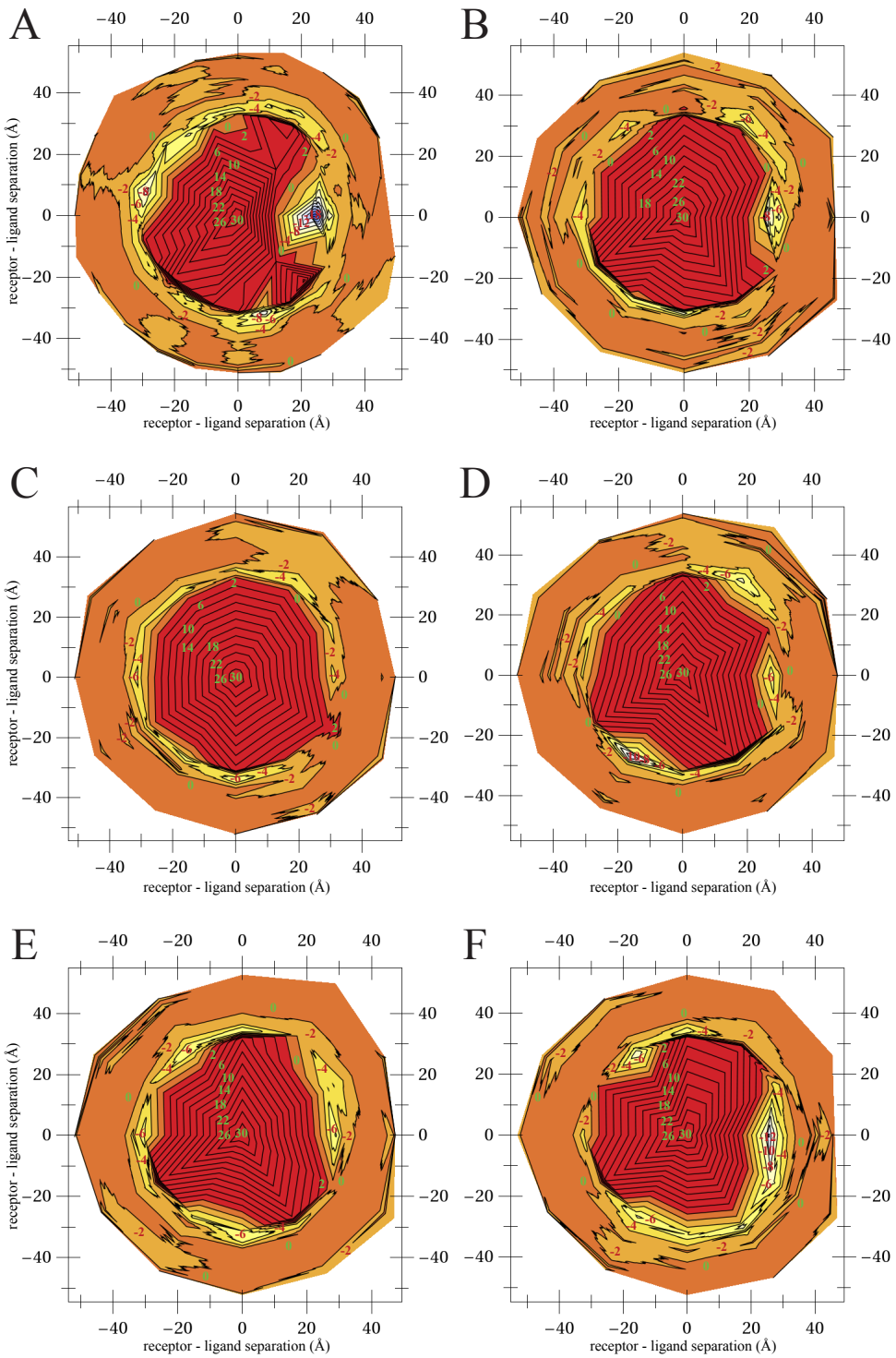

Figure 4.4: Rotations of the ligand around X: Each contour layer on the free energy surface represents 2 kcal/mol. The ligand was rotated about its own COM: A) 0°, B) 60°, C) 120°, D) 180°, E) 240°, F) 300°.

Figure 4.5 shows that the free energy surface obtained when the initial structure is at 0° (Figure 4.5 A) also has a characteristic free energy landscape different than the ones obtained when the initial structure of the ligand at position 0° was rotated on its own axis around y (Figure 4.5 B - F). The next free energy landscape with a similar shape to the one shown in Figure 4.5 A corresponds to the free energy surface when the initial structure is rotated 60° and 300° (-60°) (Figure 4.5 F) but with a less deep global minimum, ~-10 kcal / mol and ~ -8 kcal / mol, respectively. This result suggests that the ligand can rotate 60° and -60°, and still bind to the receptor.

The free energy profiles obtained when the initial structure was rotated 120°, 180° and 270° (Figure 4.4 B-E, respectively) only shows a series of local minima along of similar depth.

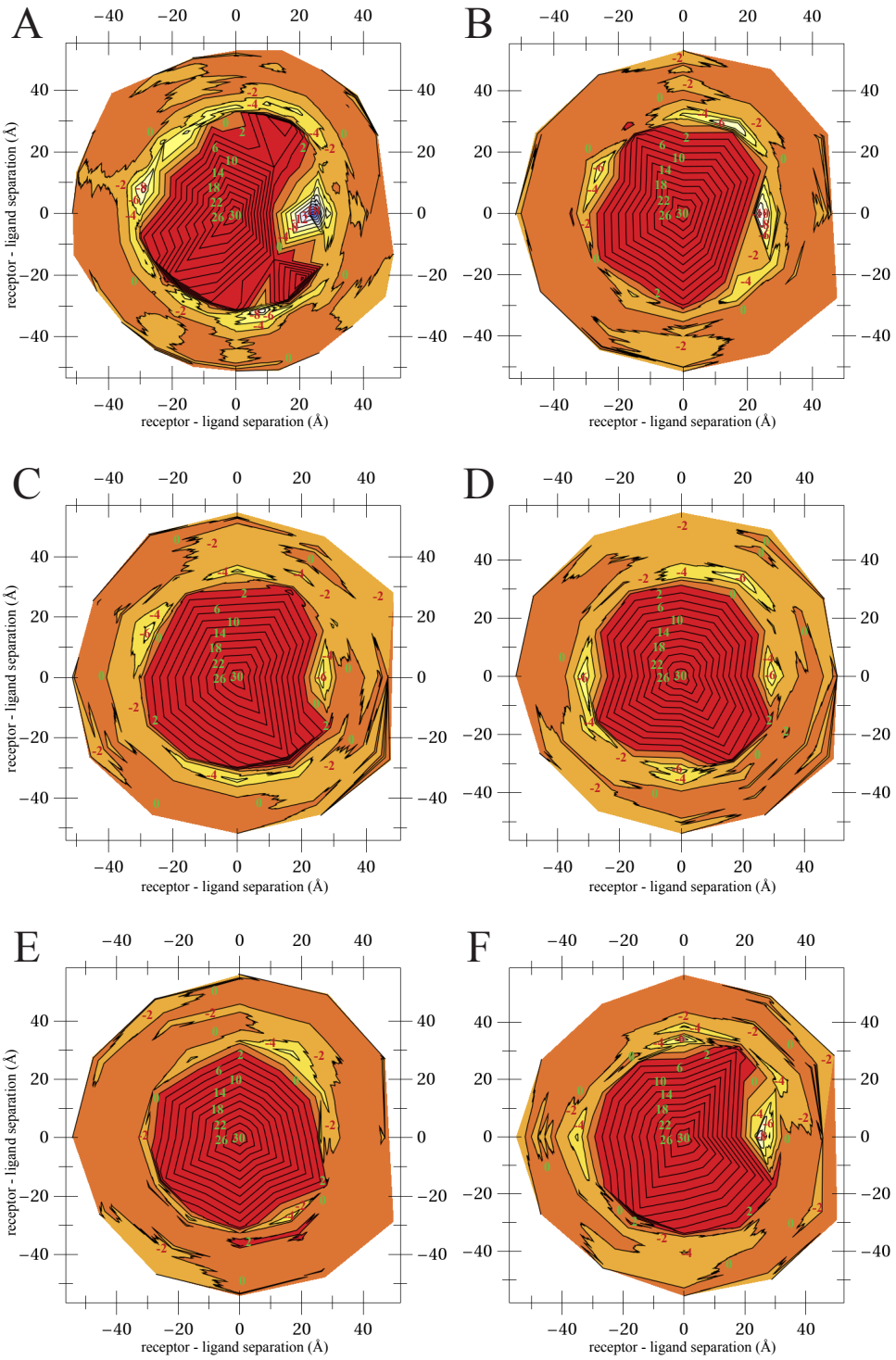

Figure 4.5: Rotations of the ligand around Y: Each contour layer on the free energy surface represents 2 kcal/mol. The ligand was rotated about its own COM: A) 0°, B) 60°, C) 120°, D) 180°, E) 240°, F) 300°.

4.3 How does the solvent affect the shape of the free energy profile?

This part focuses on finding the effect of the solvent and how this effect may contribute to the stability of the formation of the encounter and final complexes.

4.3.1.- Repulsive protein – water interactions stabilize the complex by affecting the desolvation process.

We have modified the default levels of Lennard-Jones interactions of the modified Martini 2.1 force field (martini 0). The increment of one level of interaction to all the water-protein interactions is called martini+1, for example in martini 0 the particle P4 (protein) with particle A1 (water) have an attractive interaction now the interaction will be supra-attractive. Decreasing the interactions by one level, it is called martini-1, i.e. before particle P4 (protein) with particle A1 (water) have an attractive interaction now they will be almost attractive. The same logic is applied for martini+2 and martini-2. Two new levels of interaction were created: ultra-attractive with an epsilon value of 6.0 and mega-attractive with an epsilon value of 6.5.

The results show that for all three complexes: the Barnase / Barstar complex (Figure 4.6 A), the RNase / Barstar complex (Figure 4.6 B), and the ubiquitin / ubiquitin ligase complex (Figure 4.6 C), the more repulsive the interaction between water and protein is, the more the desolvation barrier tends to disappear, making the PMF curve a very smooth downhill process. This might suggest that in the most repulsive case,

because the water tries to be away from the surface, the proteins are already in the region of the final complex just before the desolvation barrier, decreasing its size; with each member already aligned towards each other, and the interface is mostly desolvated, accelerating and stabilizing the association and decreasing the free energy minimum. The minimum is shifted to the left on the x-axis, because the repulsive water is trying to stay away from the interface, leaving the two proteins interacting closer and stronger, making the free energy minimum more negative.

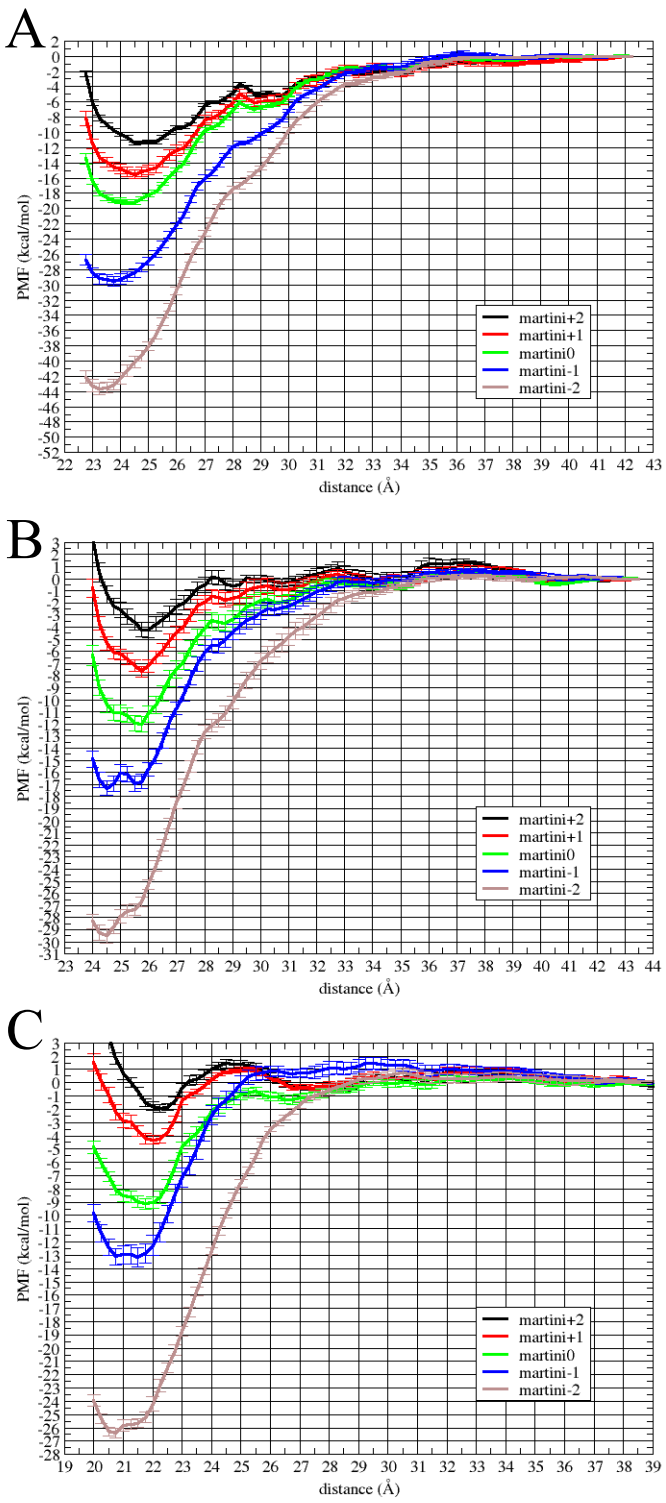

Figure 4.6: Changing the Lennard-Jones levels of interactions between the proteins and the solvent: A) Barnase / Barstar, B) RNase / Barstar, C) Ubiquitin / Ubiquitin ligase.

4.3.2. The use of a small water destabilizes the formation of the complex by increasing the area of the desolvation barrier.

This conclusion has been a very useful tool to better understand the results obtained in section 5.4.3.

All the steps and information of how the small water was designed are explained in section 2.2.2.2. Figure 4.7 shows the comparison of the free energy profiles of the Barnase / Barstar complex obtained by using the regular water (black) and the small water (red). The desolvation barrier is increased and the free energy minimum value becomes higher. The desolvation barrier might indicate that the small water beads are able to enter inside the interface, affecting the desolvation of the encounter complex. The increment of the free energy value at the minimum of the well suggests that the formation of the final complex is not favored by the presence of the small water, which destabilizes the final complex.

The difference of the free energy values obtained by using regular water minus the one obtained using the small water is depicted in green; it shows that the use of a small water does not favor the complexation process at all. The Barnase / Barstar complex presents a dominant electrostatic steering with strong long-range electrostatic interactions (Camacho, Weng et al. 1999) and high charge complementarity, the use of a small water will disrupt this charge complementarity by disrupting the electrostatic interactions i.e. salt bridges.

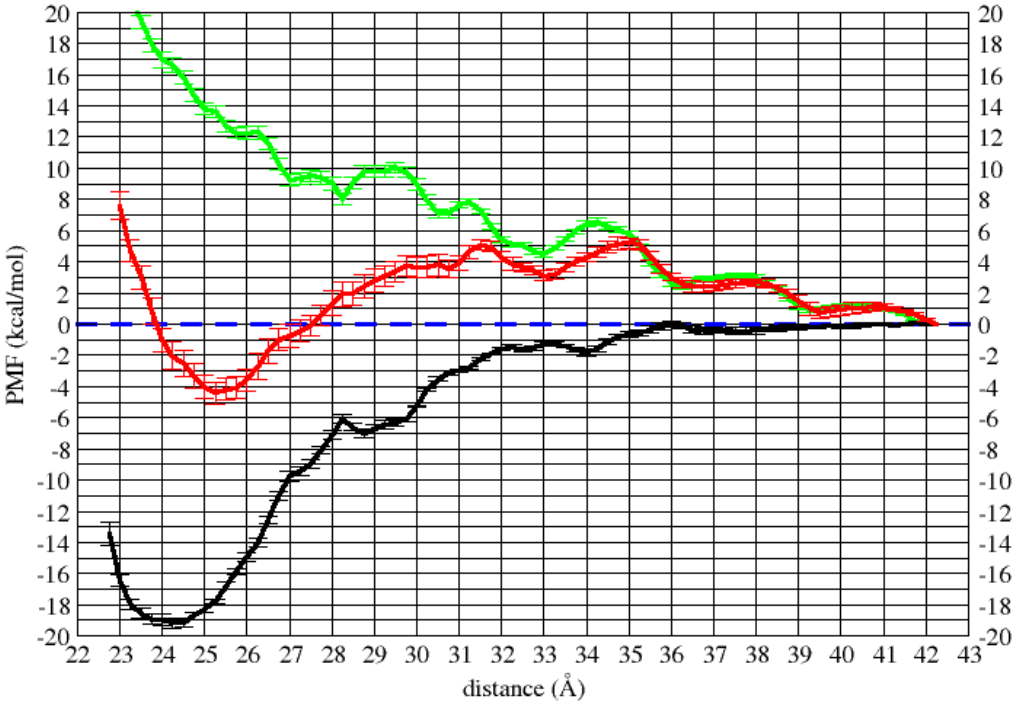

Figure 4.7: The small-water effect for the Barnase / Barstar complex: Regular water is depicted in black, small water is depicted in red and the difference between them is in green.

Figure 4.8 details the comparison of the free energy profiles of the RNase / Barstar complex obtained by using the regular water (black) and the small water (red). The desolvation barrier is also increased and the free energy minimum value becomes higher. The small water also destabilizes the formation of the final complex. The difference of the free energy values obtain by using regular water subtracting the one obtained using the small water is depicted in green; it shows that the use of a small water does not favor the complexation process. The difference in energy is positive, except for a small region where the green curve has free energy values less than zero. This result

might indicate that the small water favors the formation of the encounter complex better than the final complex.

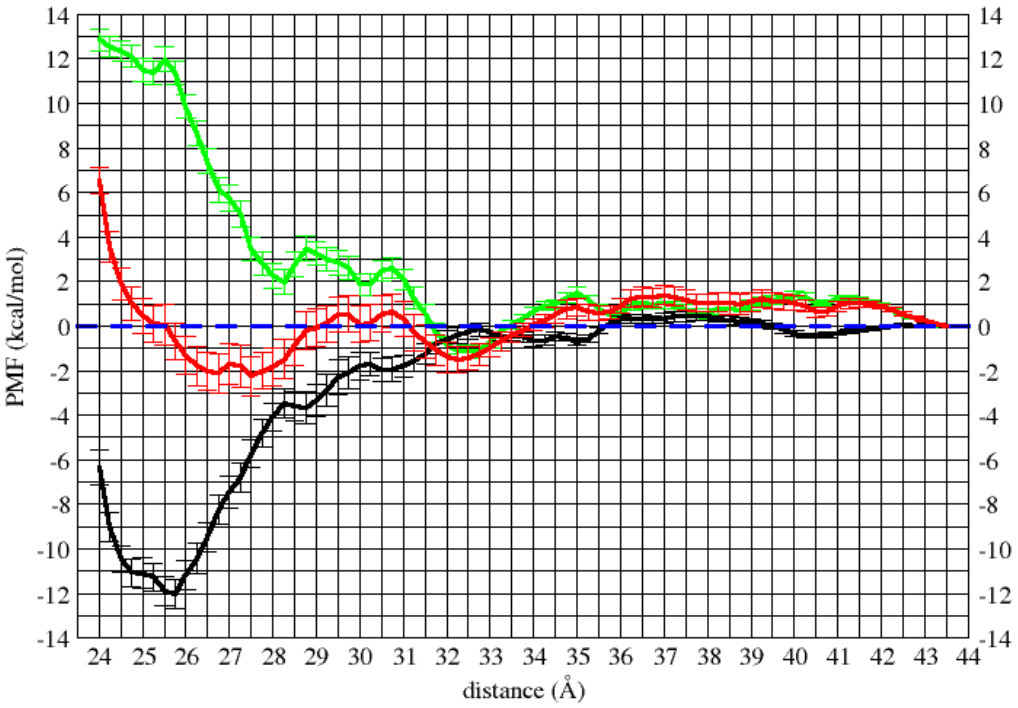

Figure 4.8: The small-water effect for the RNase / Barstar complex: Regular water is depicted in black, small water is depicted in red and the difference between them is in green.

Figure 4.9 shows the comparison of the free energy profiles of the Ubiquitin / Ubiquitin ligase complex obtained by using the regular water (black) and the small water (red). The desolvation barrier is also increased and the free energy minimum value became slightly higher by less than 2 kcal/mol. The small water does not destabilize the formation of the final complex as much as in the other two cases. The difference of the free energy values

obtained by using regular water subtracting the one obtained using the small water is depicted in green. The difference in energy is positive, except for a region between the encounter complex and the final complex, where the green curve has free energy values less than zero very close to the encounter complex. The increments of the desolvation region before the encounter complex indicate that more water molecules are between the two proteins. Just after the formation of the encounter complex there is the desolvation process; in the small water case, because the size of the bead is smaller, the desolvation seems to be favored, just after the formation of the encounter complex. When the two proteins are closer to the final complex position, some small water might be trapped in the interface, destabilizing by ~2 kcal/mol the free energy minimum.

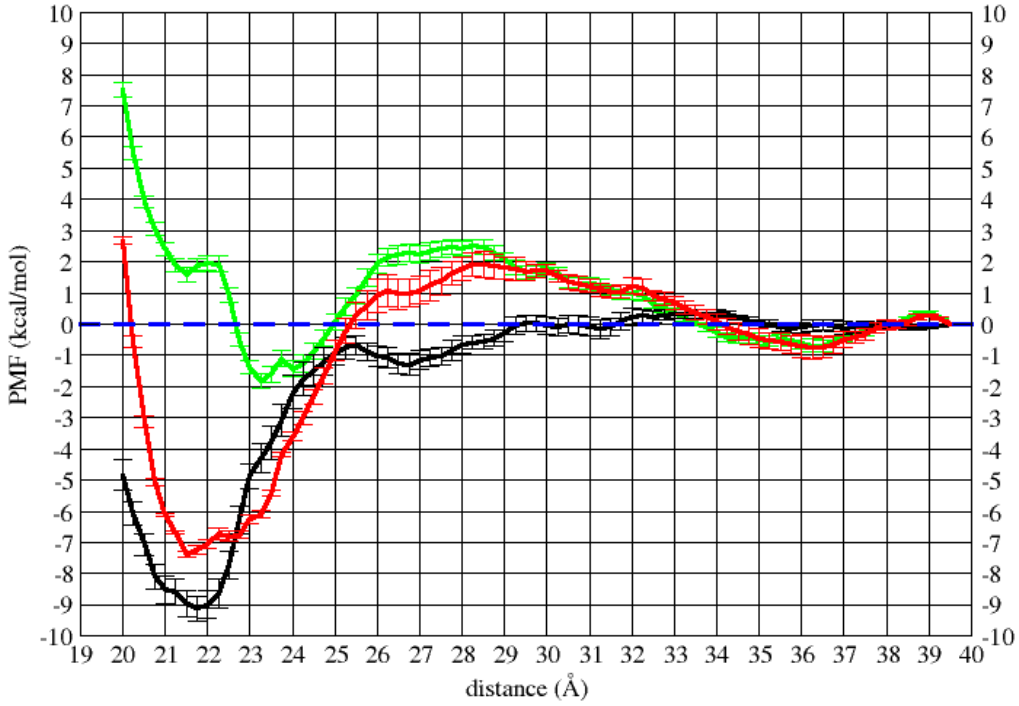

Figure 4.9: The small-water effect for the Ubiquitin / Ubiquitin ligase complex: Regular water is depicted in black, small water is depicted in red and the difference between them is in green.

5

Can ELNEDIN models be used to distinguish native protein complexes from non-native ones?

5.1 Introduction

The main objective of this chapter is to test if the free energy profile obtained using ELNEDIN-based protein models is able to distinguish native from non-native interfaces. In order to achieve this goal we will use the same three protein complexes used in chapter 3: 1) The Barnase - Barstar complex (PDB I.D.: 1BRS), 2) The RNase – Barstar complex (PDB I.D.: 1AY7), and 3) The Ubiquitin – Ubiquitin Ligase complex (PDB I.D.: 2OOB). The docking program ZDOCK will generate non-native poses, which will be filtered in order to get representative non-native configurations. The native and non-native configurations will be further simulated using the ELNEDIN approach with elastic network parameters Rc/K_{SPRING} of 1.2/1000. We will further compare and analyze the obtained free energy profiles.

5.2 Steric aspects of recognition

The protein-docking program ZDOCK was used to generate docked conformations. ZDOCK 3.0.1 (Chen and Weng 2002; Chen, Li et al. 2003), has an algorithm based on the Fast-Fourier transform (FFT) correlation approach and a scoring scheme based on pairwise shape complementarity, an electrostatic energy term and a desolvation energy term. Using the default parameters, 2000 poses were generated for each of the three selected complexes.

Only for comparison purposes, another docking-program FTDOCK (Gabb, Jackson et al. 1997) was used. FTDOCK 2.0 has an algorithm based on a modification of the Fast-Fourier transform (FFT) correlation approach, the Katchalski-Katzir algorithm (Katchalski-Katzir, Shariv et al. 1992). The scoring scheme of FTDOCK is based primarily on a shape complementarity score and an electrostatic energy term. Using the default parameters, 2000 poses were generated for each of the three complexes.

ZDOCK and FTDOCK are initial-stage protein-docking programs; for that reason, it is not expected that the native or native-like pose would be ranked first or inside the first 10 outputs. The 2000 output poses were picked for each program. In order to filter the 2000 configurations, for each pose the buried surface area (BSA) and the shape complementarity (SC) was computed.

The buried surface area was calculated in the following way:

$$\text{BSA} = \text{SASA of ligand} + \text{SASA of receptor} - \text{SASA of complex} \qquad (5.1)$$

Where the solvent accessible surface area (SASA) was computed using NACCESS 2.1 (Hubbard and Thornton 1993) with the default probe size of 1.4 Å.

The shape complementarity (SC) was calculated using the algorithm proposed by (Lawrence and Colman 1993) with a probe size of 1.7 Å and other default parameters.

Figure 5.1 shows the SC values plotted versus the BSA values for Barnase / Barstar complex (A and B), the RNaseA / Barstar complex (C and D), and Ubiquitin / Ubiquitin ligase complex (E and F) using the two docking algorithms ZDOCK (left column) and FTDOCK (right column). Each dot in the figure represents a pose generated by ZDOCK or FTDOCK. Notice the position of the native structure (red circle). The native structures have higher values of SC with respect to the predicted poses. This result shows the importance of the SC as one of the main sterical aspects of recognition. SC is considered a representative of the 'lock-and-key' mechanism of protein binding (Papoian and Wolynes 2003). SC is also the main criterion to rank predicted complexes for many docking programs because it is a good indicator of geometric matching between proteins. It was demonstrated that SC values for complexes vary from 0.76 to 0.70 for oligomeric or protein/protein inhibitor complexes and from 0.68 to 0.64 for antibody/antigen interfaces, the latter being the one with significantly poorer shape complementarity (Lawrence and Colman 1993).

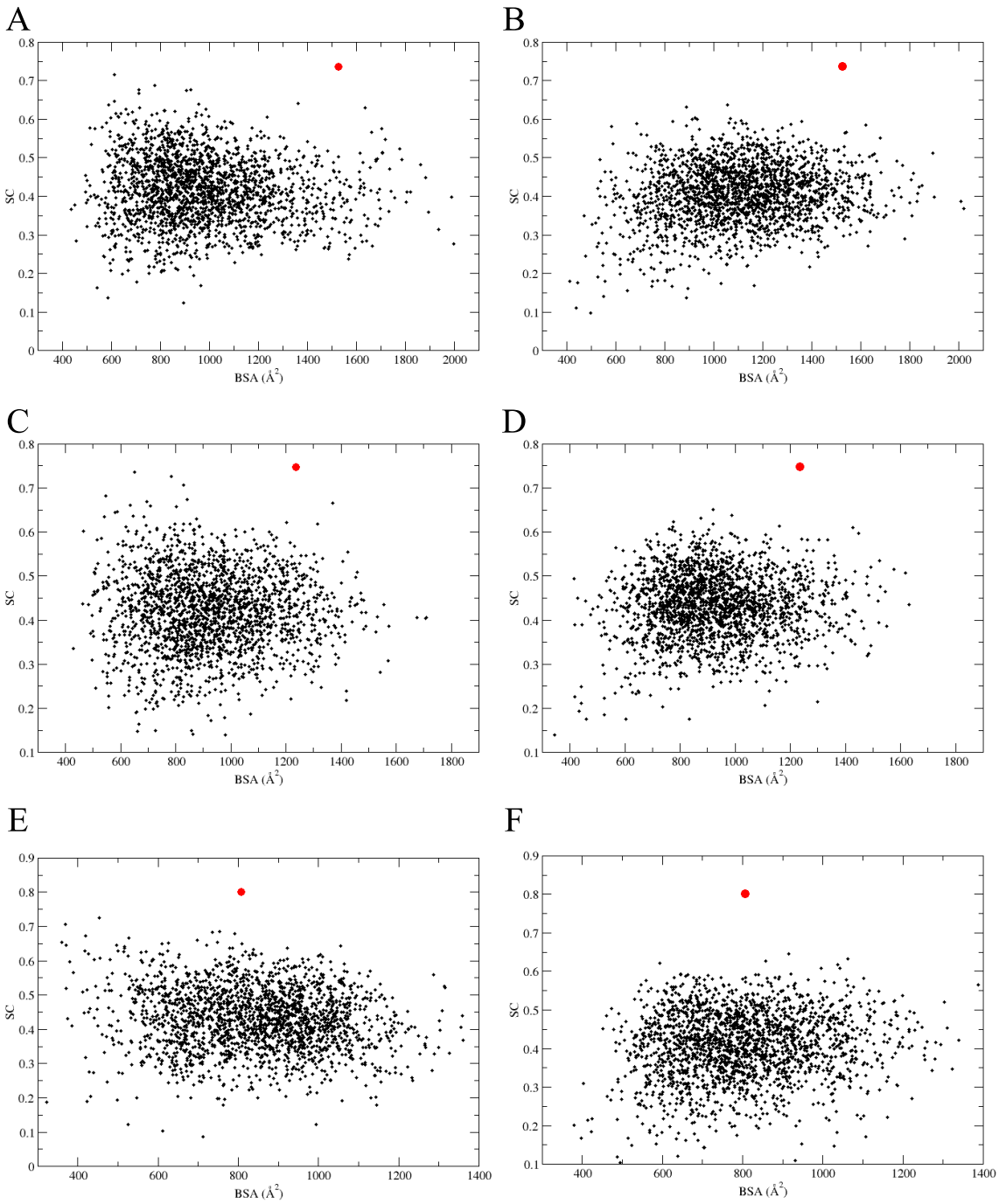

Figure 5.1: Comparison of SC vs. BSA plots obtained by ZDOCK (left) **and FTDOCK** (right)**:** Notice the position of the native structure (red circle). A) Barnase / Barstar: ZDOCK, B) Barnase / Barstar: FTDOCK, C) RNase / Barnase: ZDOCK, D) RNase / Barnase: FTDOCK, E) Ubiquitin / Ubiquitin ligase: ZDOCK, F) Ubiquitin / Ubiquitin ligase: FTDOCK.

Figure 5.1 also shows that there is not a preferable position of the native conformation with respect to the non-native structures in terms of BSA. The overall comparison between the plots obtained by using ZDOCK and FTDOCK look very similar; there is not a perceivable difference between them. Arbitrarily, we decided to perform our calculation using only ZDOCK.

To evaluate if for other complexes the native conformation has the same position preference of having a relative high SC value with respect to the non-native complex, 43 complexes were chosen from the "rigid-body set" of protein complexes of the Protein Docking Database Benchmark 4.0 (Hwang, Vreven et al. 2010). The "rigid-body set" is made by proteins that do not suffer big conformational changes upon complexation. See section 2.2.1.1 for details on how those complexes were chosen. The SC versus BSA values are plotted for each complex (Figures 5.2 - 5.6). In Figures 5.2 to 5.6, it is observed that the position of the native structure is above the predicted poses, corroborating that the native structure should have a high SC value. The native-like conformation was obtained using the root mean square deviation (RMSD) (Equation 2.21) between the ligand of the predicted conformation and the ligand of the crystal structure of the complex (native), the one with the lowest RMSD value should correspond to the native-like conformation. The native-like conformation falls very close to the native conformation with respect to the BSA value.

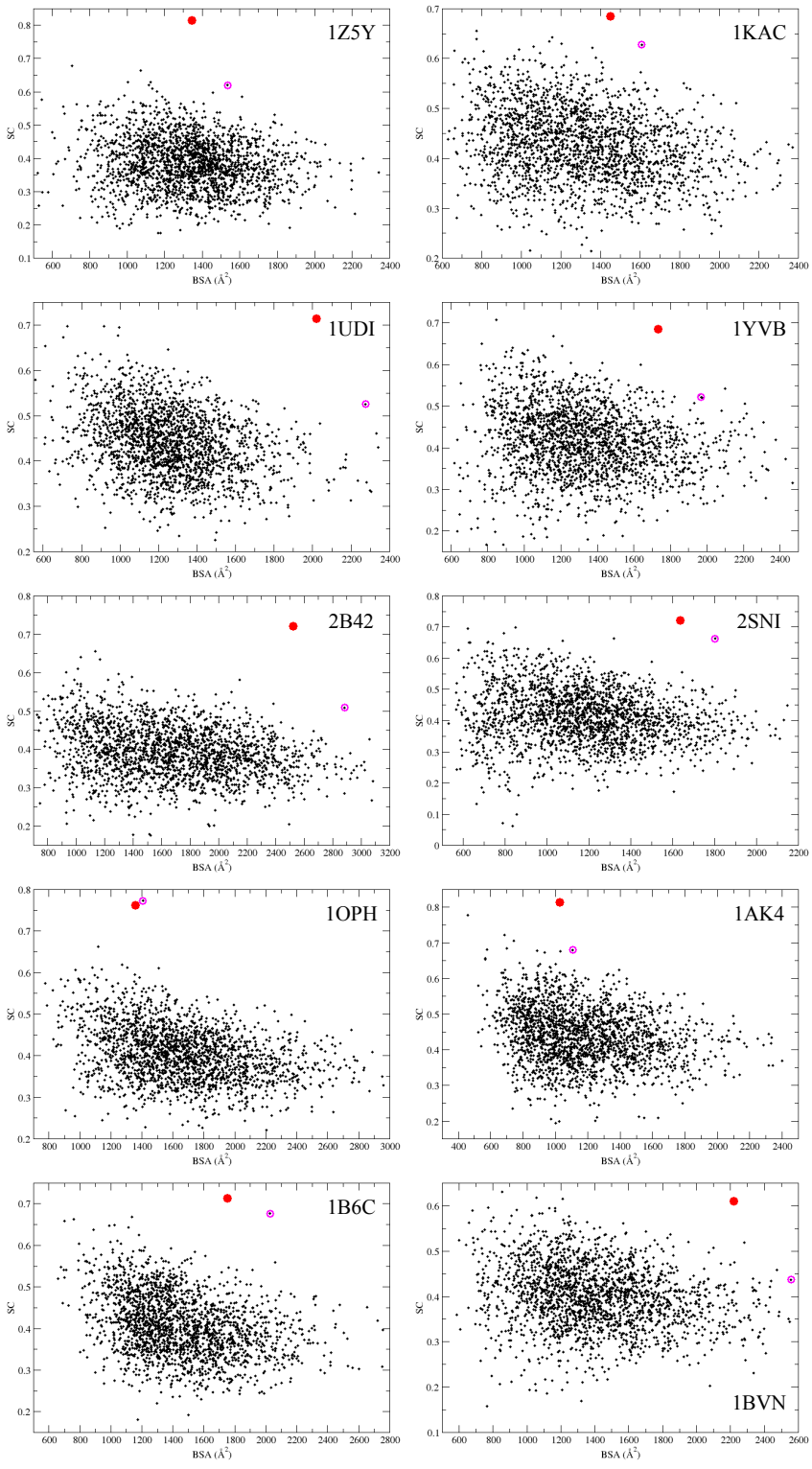

Figure 5.2: SC vs BSA Part 1: The native structure is represented by a red circle, and the native-like structure is enclosed in a magenta circle.

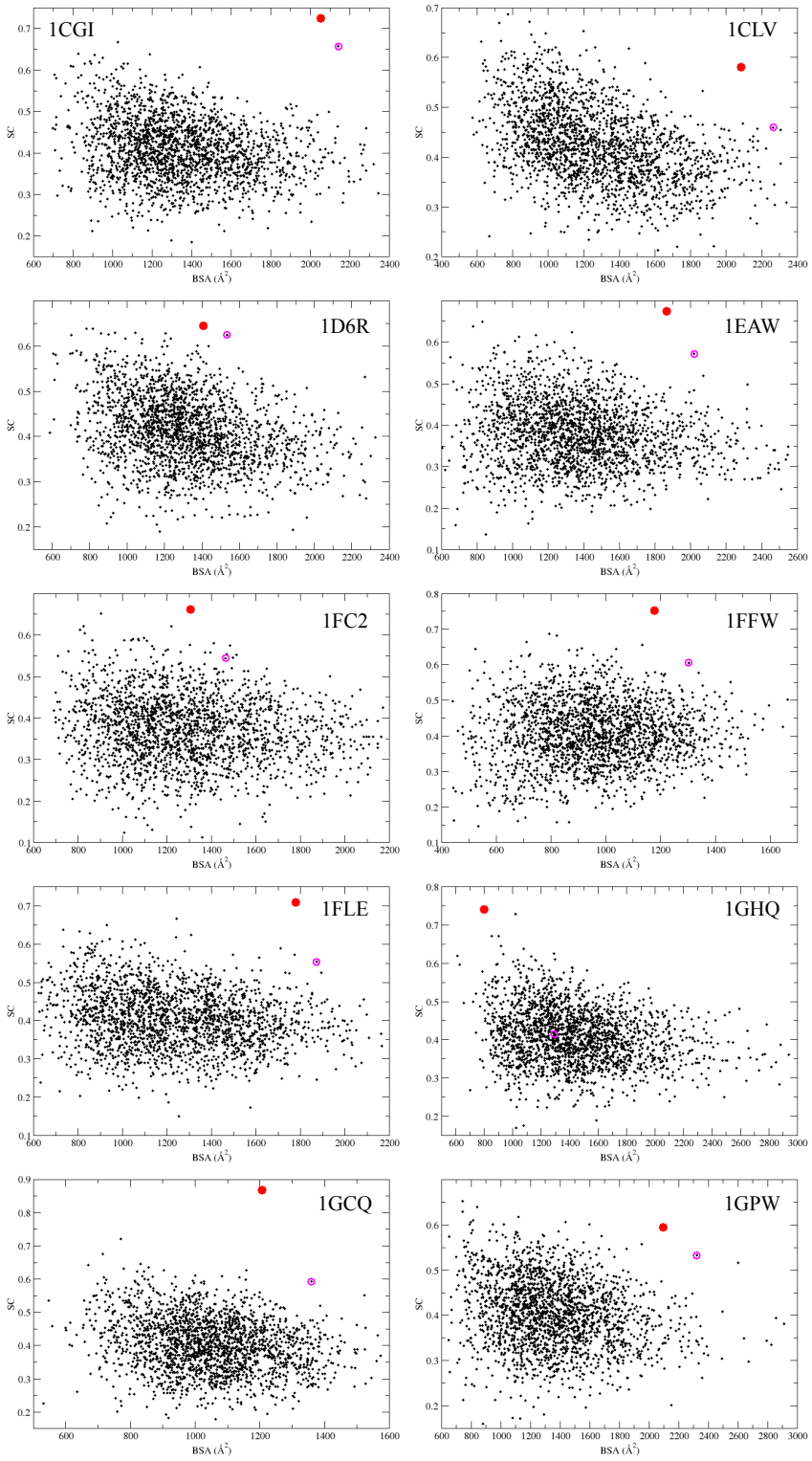

Figure 5.3: SC vs BSA Part 2: The native structure is represented by a red circle, and the native-like structure is enclosed in a magenta circle.

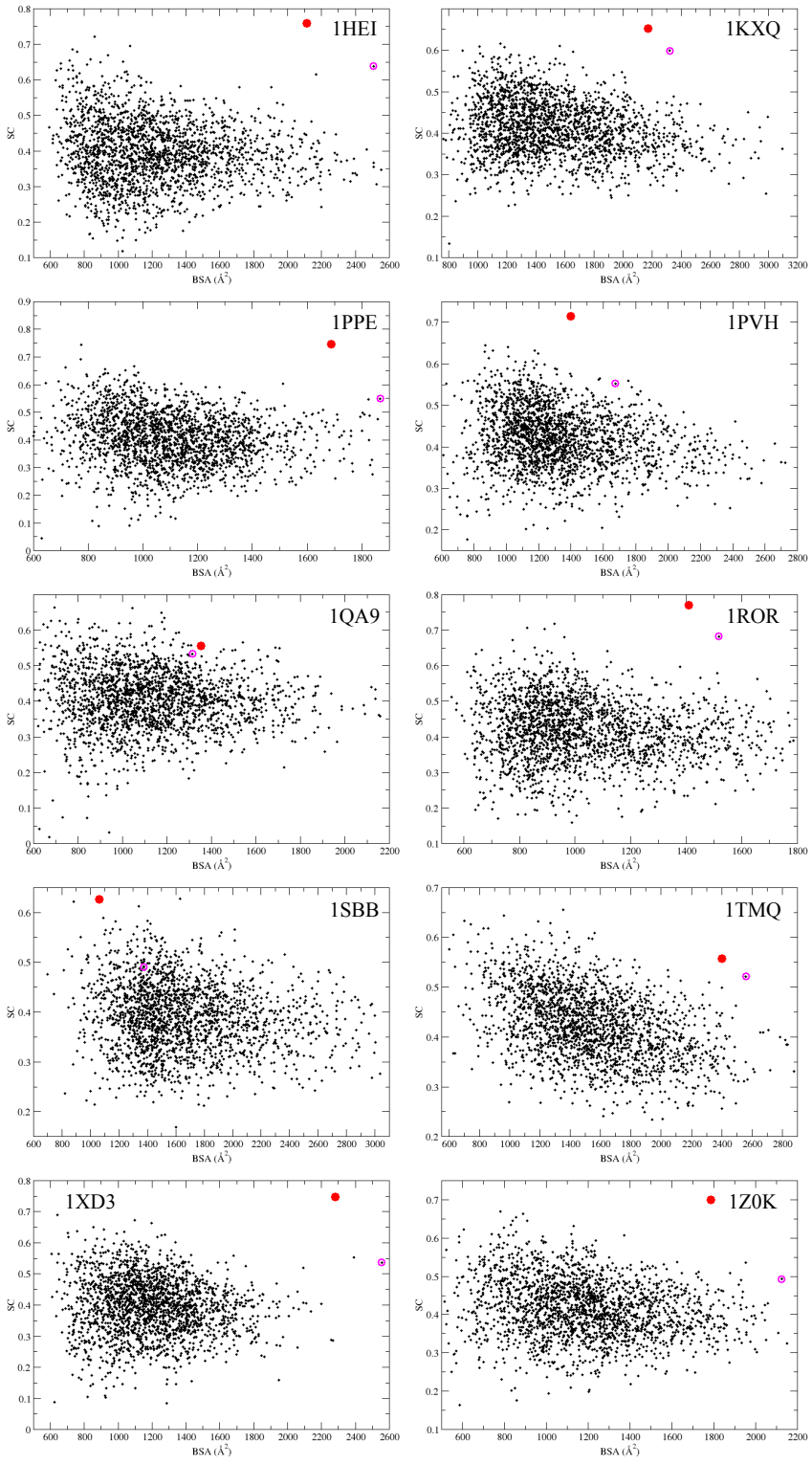

Figure 5.4: SC vs BSA Part 3: The native structure is represented by a red circle, and the native-like structure is enclosed in a magenta circle.

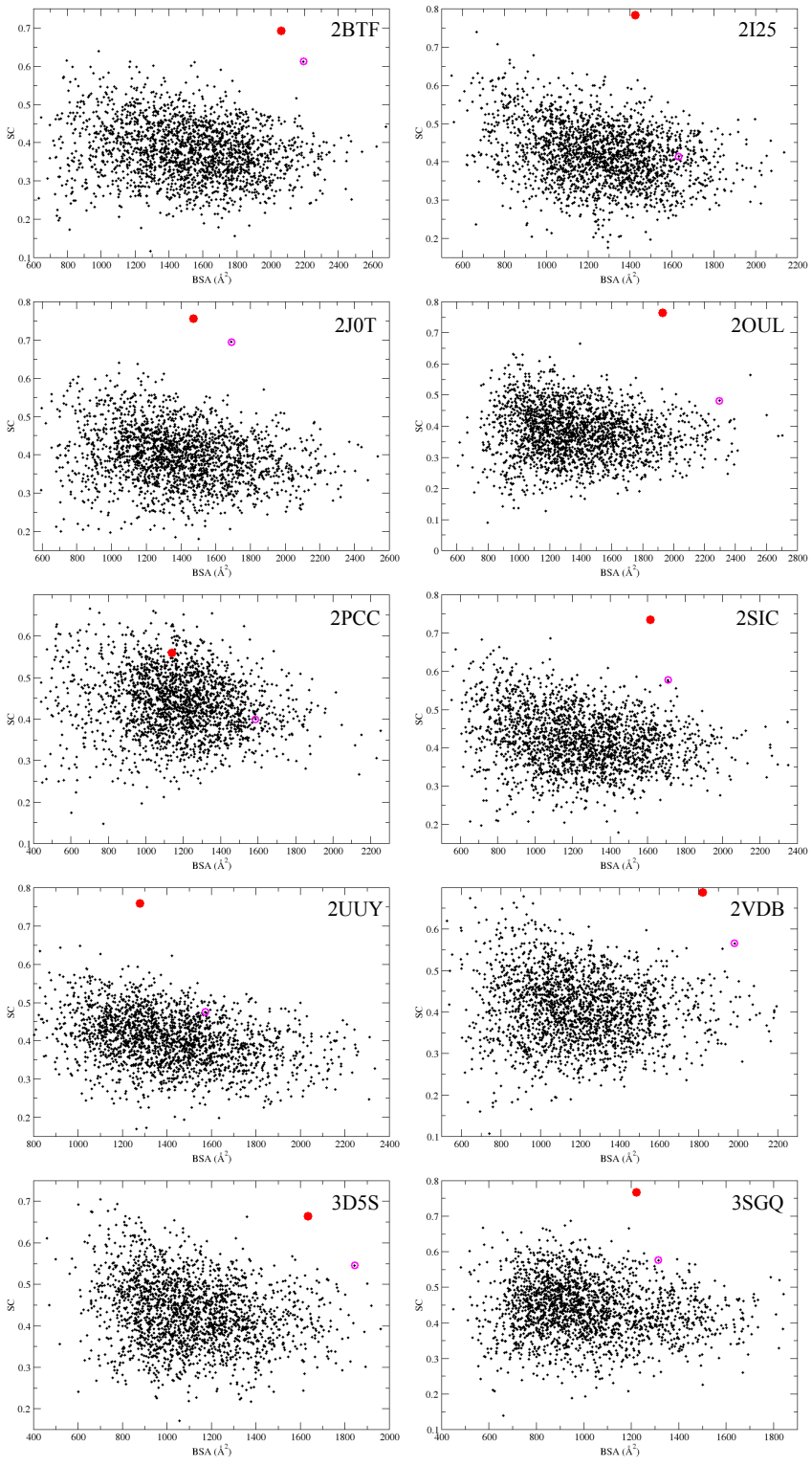

Figure 5.5: SC vs BSA Part 4: The native structure is represented by a red circle, and the native-like structure is enclosed in a magenta circle.

In all of the cases the native pose is in the upper part of the plot, and in almost all of the cases the native-like pose is also in the upper region of the plot so we decided to pick the poses that are in the upper part of the cloud made by the predicted poses (Figure 5.6 green). To pick the chosen poses (green) the BSA axis of each plot was subdivided in bins of 200 Å with a sliding window of 25 Å. To each window, only the poses with high SC values (the ones that are higher than the average plus one standard deviation) were picked. Also, upper outliers that satisfy the condition to be greater or equal to three standard deviations were chosen.

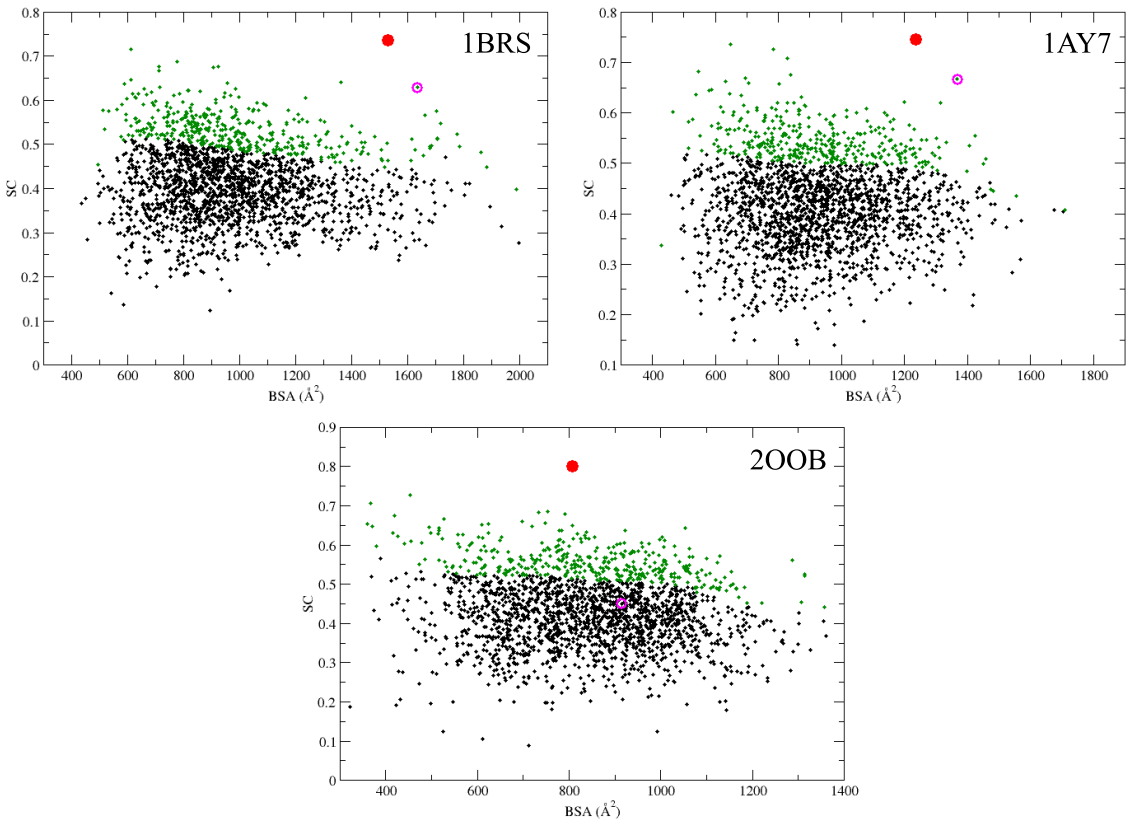

Figure 5.6: SC vs BSA Part 5: The native structure is represented by a red circle, and the native-like structure is enclosed in a magenta circle. The chosen structures are depicted in green.

5.3 Searching for the promiscuous surface

The chosen poses (Figure 5.6 green) obtained in step 5.2 need to be filtered because it won't be possible to test for each system ~ 400 picked poses (Figure 5.6 green) in a reasonable amount of time. We have developed a representation, where the COM of the receptor is in the middle box (box 0), and the COM of the ligand might be in any of the other 26 boxes surrounding the middle box (Figure 5.7).

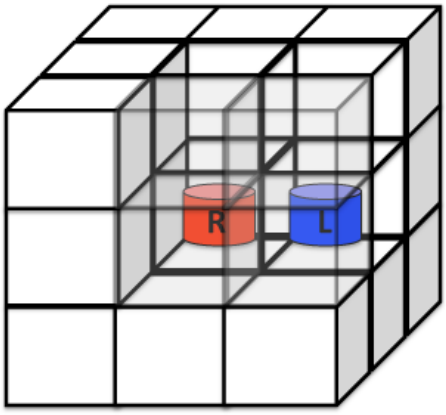

Figure 5.7: Boxing the picked poses: Cartoon representation of how the ligand (blue) was distributed around the receptor (red).

We picked only the most populated boxes. A box will be chosen if it encloses more than 20 poses (Figure 5.8). This number was chosen arbitrarily.

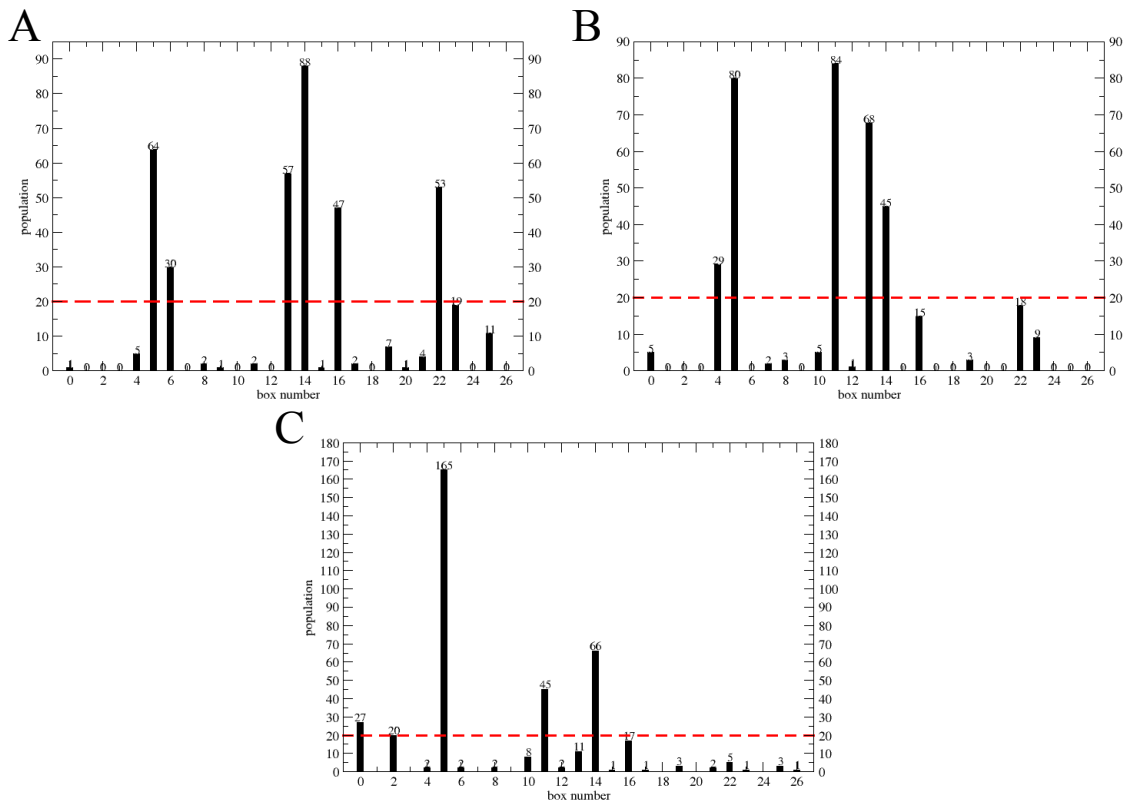

Figure 5.8: Population per box: A) Barnase / Barstar, B) RNase / Barstar, C) Ubiquitin / Ubiquitin ligase. A segmented red line marks the minimum cut-off of 20.

Inside each box there are several rotations and translations of the ligand. One way to group all the poses that are in the same orientation is by performing a hierarchical clustering. To the poses inside each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. The RMSD calculation was performed only for the ligand Cα, we did not include the receptor in the calculation of the RMSD because it is fixed. We have chosen the clusters with the largest number of members within a fixed cluster distance less than 10 Å (Figure 5.9). For visualization purposes Figure 5.9 shows only one dendrogram, more information will be found in the appendix. A similar criterion is used by the initial-stage docking program ClusPro (Comeau, Gatchell et al. 2004; Comeau, Gatchell et al. 2004).

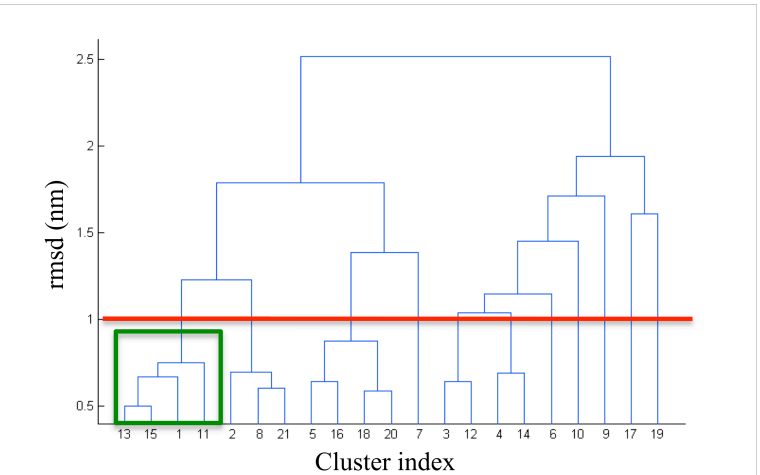

Figure 5.9: Dendrogram obtained from the hierarchical clustering of the poses inside box 6 of Barnase / Barstar: A red line is marking the fixed cluster distance of 10 Å. The most populated grouped cluster is enclosed in a green rectangle.

In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. With this procedure, it is possible to pick the native-like structure in the selected representative poses in almost all cases, except for ubiquitin – ubiquitin ligase complex (2OOB) where ZDOCK does not find a native-like pose, the closest pose to the native has an RMSD value of 3 Å away with respect to the native structure with a shape complementarity value of 0.45, a very low value in comparison with the native which have a SC value of 0.81.

5.4 Finding the needle in the haystack

To test if the ELNEDIN approach is able to distinguish between the native and non-native poses, potential of mean force calculations are performed on the chosen representative picked poses of Barnase / Barstar complex (see methods 2.2.1.2.1), the RNaseA / Barstar complex (see methods 2.2.1.2.2), and the Ubiquitin / Ubiquitin ligase complex (see method 2.2.1.2.3).

As a preliminary result for future studies, we also test the ability of ELNEDIN in distinguishing native and non-native poses for a pair of proteins that suffer large conformational changes upon complexation; this one complex belongs to the "difficult cases" from the protein-protein docking benchmark (Chen, Mintseris et al. 2003; Hwang, Pierce et al. 2008; Hwang, Vreven et al. 2010): the bound-bound and bound-unbound form of the NucA / NuiA complex (see methods 2.2.1.2.4).

5.4.1 The Barnase / Barstar case:

The interaction between Barnase and Barstar is one of the strongest and fastest interactions between proteins. This interaction needs to be fast and strong to prevent cell death. Its complexation is principally driven by electrostatics and its association rate of 10^8 $s^{-1}M^{-1}$ (Buckle, Schreiber et al. 1994) is one of the fastest known in biological systems (Hoefling and Gottschalk 2010). The binding interface consists mainly of polar and charged residues, and shows a high shape and electrostatic complementarity.

After performing the steps explained in section 5.3 (a detailed description of each step is given in appendix A.1), the following poses were chosen and named after the location of the box in which they were found: box 5, box 6, box 13, box 14 (this pose corresponds to the native-like pose), box 16 and box 22 (Figure 5.10).

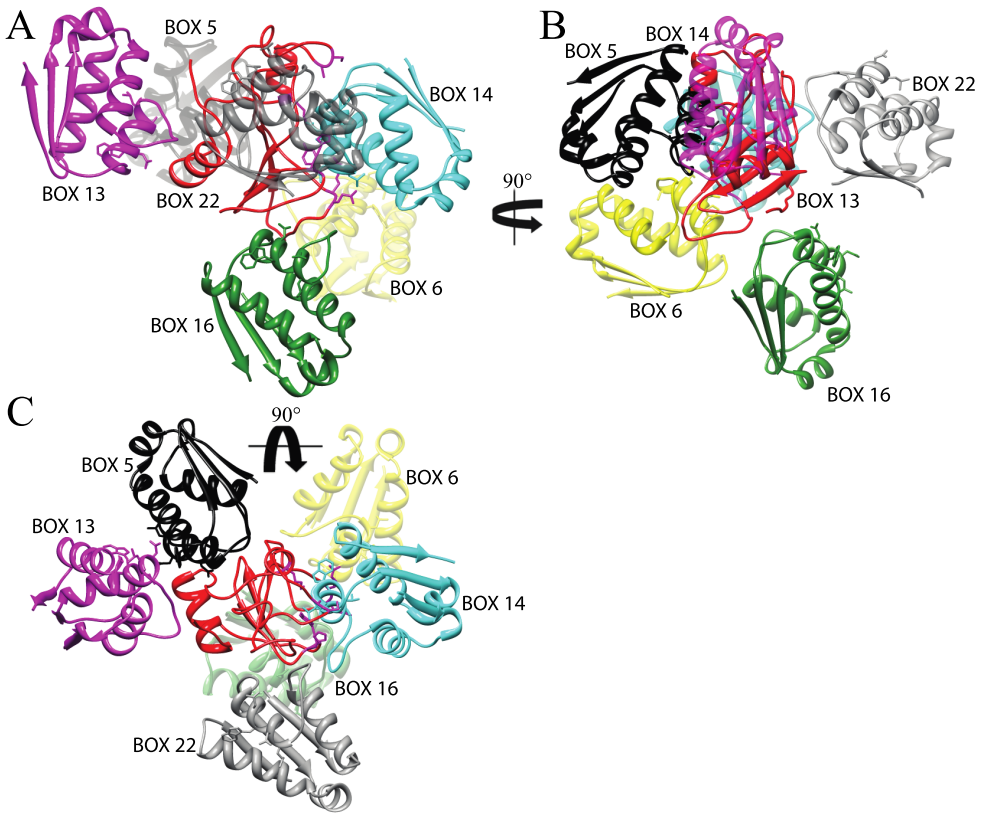

Figure 5.10: Selected representative poses for the Barnase / Barstar complex.

For each pose, the respective values of BSA, SC and RMSD with respect to the native ligand are shown in Table 5.1.

Table 5.1: Barnase - Barstar representative poses: List of the representative poses of Barnase - Barstar complex with their respective BSA, SC and RMSD value versus the ligand native conformation.

	BSA (Å^2)	SC	RMSD (Å)
NATIVE	1530	0.74	0.0
BOX5	1230	0.50	43.6
BOX6	929	0.56	31.5
BOX13	743	0.59	55.9
BOX14*	1636	0.62	1.4
BOX16	808	0.54	39.4
BOX22	977	0.54	34.9

The ELNEDIN approach was done using the rigid elastic network of R_C = 1.2 nm and k_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} (1.2/1000). All the steps of how the PMF were obtained are explained in sections 2.2.4 and 2.2.5. The native structure has the highest SC value of 0.74 and BSA of 1530 Å^2. The native-like pose has the second highest SC value of 0.62 with BSA of 1636 Å^2. The free energy profile (Figure 5.11 D) obtained for the native-like structure is very similar to the one obtained by the native conformation (black), with a small ~2 kcal/mol barrier ~ 4 Å away from the minimum of the well, and a minimum free energy value no more than 2 kcal/mol of difference from the one obtained by native conformation (Figure 5.11-D); indicating that the ELNEDIN approach was able to identify the native-like conformation. The free energy profile for the other boxes shows different shapes and lower binding free energy with respect to the free energy profile of the native or native-like pose.

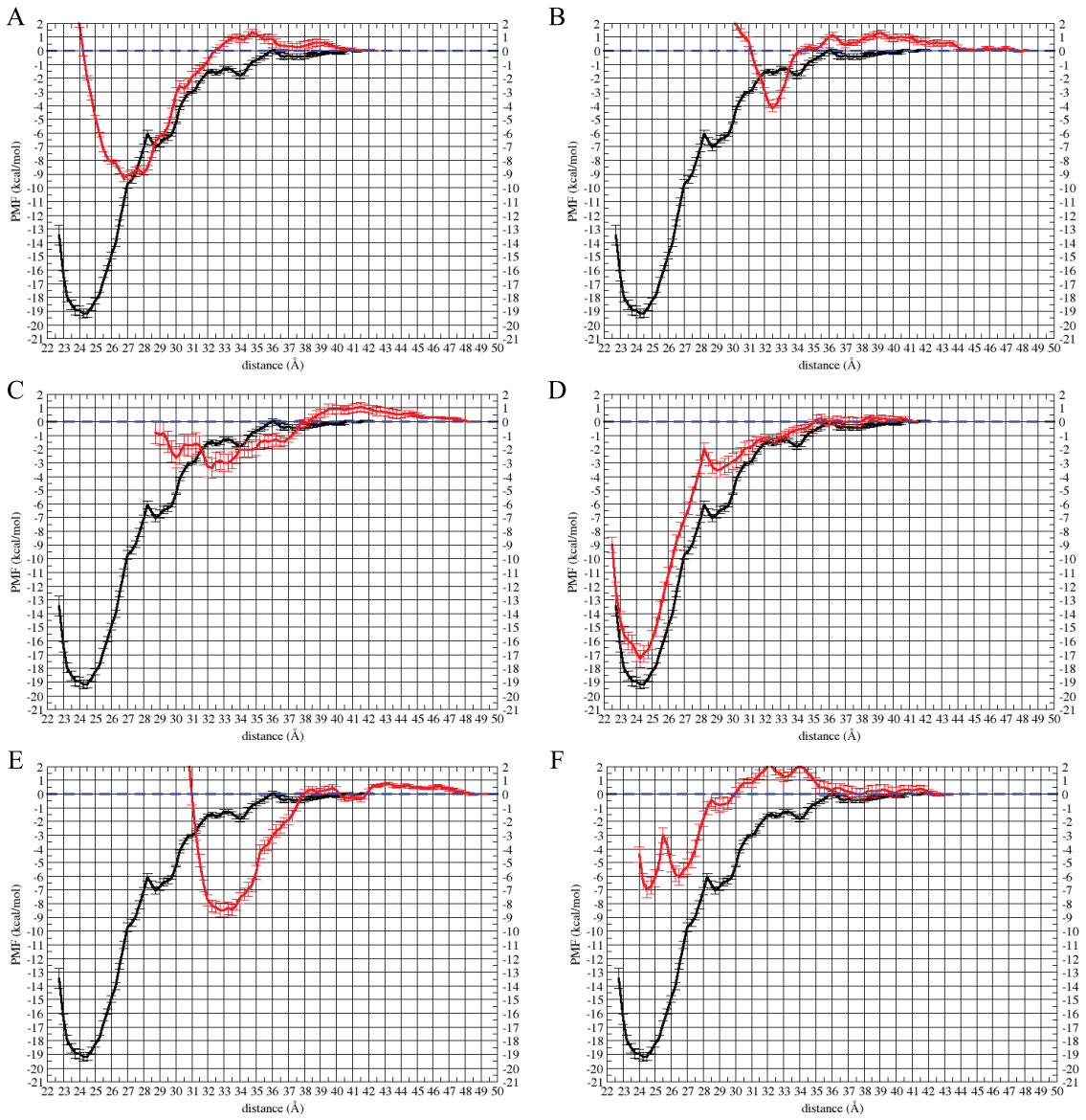

Figure 5.11: Free energy profile comparison of the dissociation of the Barnase / Barstar chosen poses (red) versus the native conformation (black): (A) Box 5, (B) Box 6, (C) Box 13, (D) Box 14 or 'native-like', (E) Box 16, (F) Box 22.

5.4.2 The RNase / Barstar case:

The complexation of the RNase / Barstar complex is driven by electrostatics. The interaction between the protein members is also strong and fast but less than the Barnase / Barstar complex. The binding interface consists mainly of polar and charged residues, and shows a high shape and electrostatic complementarity.

After performing the steps explained in the section 5.3 (a detailed description of each step is given in appendix A.2), the following poses were chosen and named after the location of the box in which they were found: box 4, box 5 (this pose corresponds to the native-like pose), box 11, box 13 and box 14 (Figure 5.12).

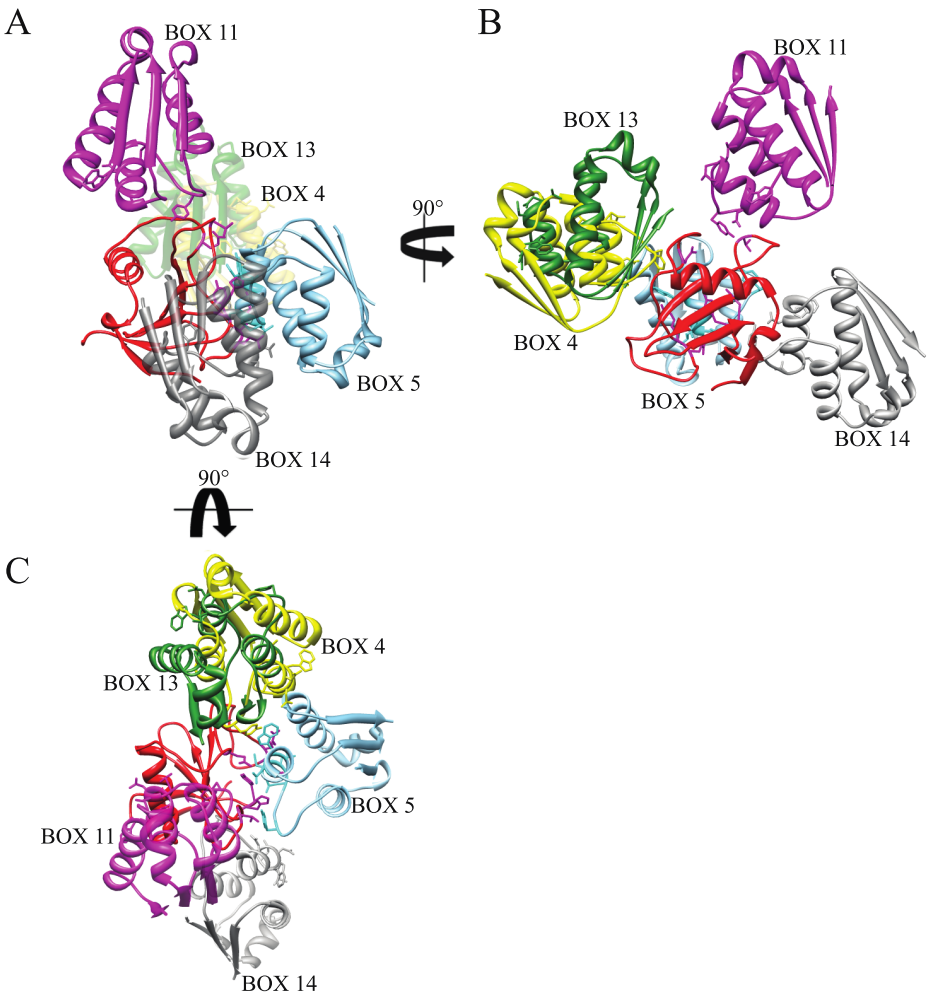

Figure 5.12: Selected representative poses for the RNase / Barstar complex.

For each pose, the respective values of BSA, SC and RMSD with respect to the native ligand are shown in Table 5.2.

Table 5.2: RNase - Barstar chosen poses: List of the chosen poses of RNase / Barstar complex with their respective BSA, SC and RMSD value versus the ligand native conformation.

	BSA (Å^2)	SC	RMSD (Å)
NATIVE	1237	0.75	0.0
BOX4	791	0.51	36.1
BOX5*	1369	0.67	2.1
BOX11	705	0.60	45.6
BOX13	1129	0.51	36.5
BOX14	840	0.52	37.1

The ELNEDIN approach was done using the rigid elastic network of R_C = 1.2 nm and k_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} (1.2/1000). All the steps of how the PMF were obtained are explained in sections 2.2.4 and 2.2.5.

The native structure has the highest SC value of 0.75 and BSA of 1237 Å^2. The resulting free energy profile (Figure 5.13) for each pose shows that box 4 (Figure 5.13-A) and the native-like pose (box 5) (Figure 5.13-B) have both the lowest free energy value after the native, but the position of the minimum of the well in box 4 is ~4 Å away from the native or native-like. Box 4 has a very low SC value of 0.51 with a low BSA of BSA of 791 Å^2. On the other hand, the SC and BSA value of box 5 is very close to the native, 0.67 and 1369, respectively. The native-like or box 5 pose has a free energy profile similar to the free energy profile of the native structure (black), with a ~1 kcal/mol barrier at ~ 3 Å away from the minimum of the well and a minimum free energy difference no more than 1 kcal/mol from the native conformation (Figure 5.13-B); indicating that the ELNEDIN approach was able to identify the native-like conformation. The free

energy profile for the other boxes shows different shapes and lower binding free energy with respect to the free energy profile of the native or native-like pose.

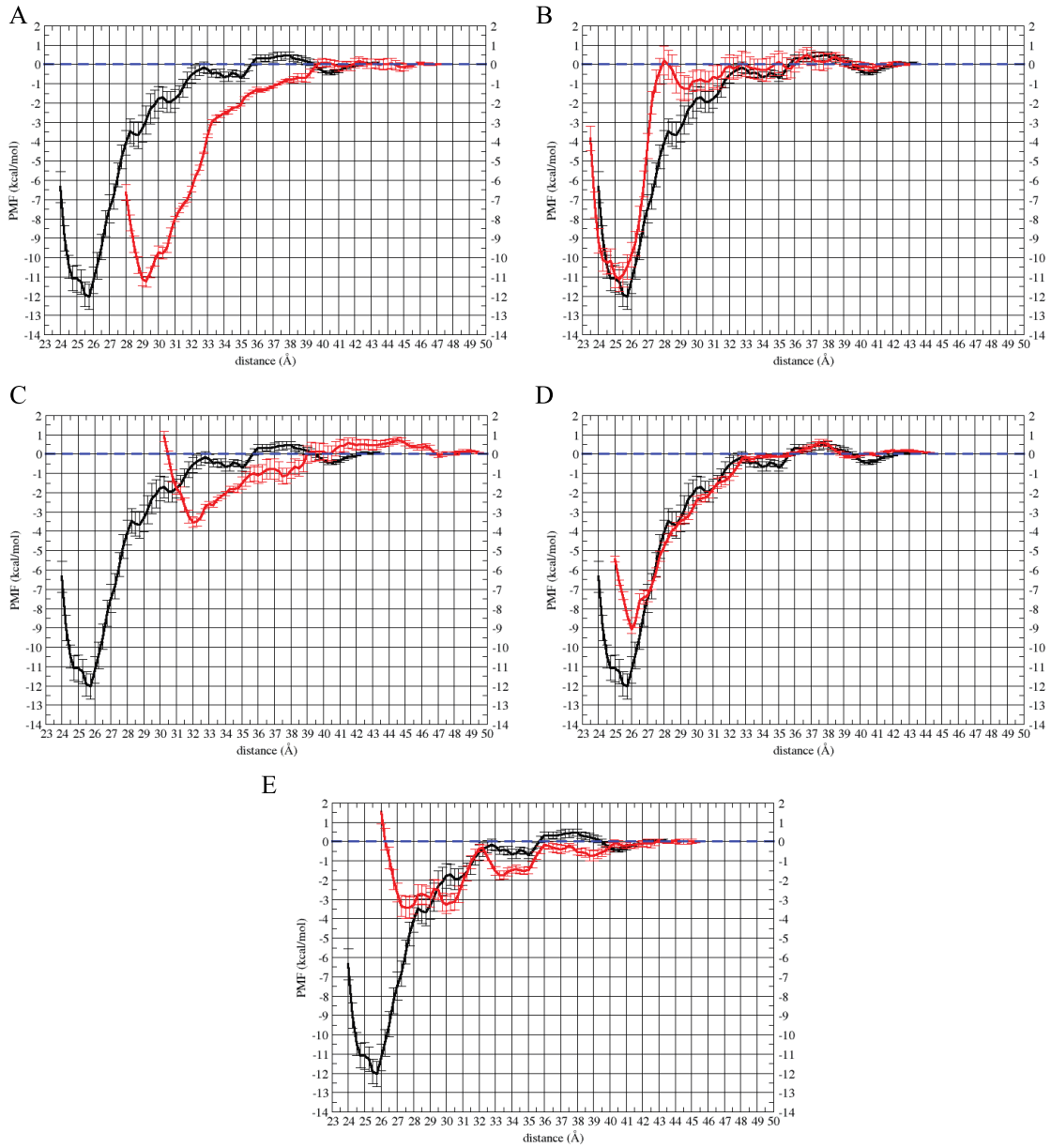

Figure 5.13: Free energy profile comparison of the dissociation of the RNase / Barstar chosen poses (red) versus the native conformation (black): (A) Box 4, (B) Box 5 or 'native-like', (C) Box 11, (D) Box 13, (E) Box 14.

5.4.3 The Ubiquitin / Ubiquitin ligase case:

The Barnase / Barstar complex and the RNase / Barstar complex have an interface composed mainly of polar and charged residues but the interface of the Ubiquitin / Ubiquitin ligase complex is made by hydrophobic residues.

After performing the steps explained in the section 5.3 (a detailed description of each step is given in appendix A.3), the following poses were chosen and named after the location of the box in which they were found: box 0, box 2, box 5 (this pose corresponds to the native-like pose), box 11 and box 14 (Figure 5.14). Box 5 is the native-like pose but it is not the structure with the lowest RMSD with respect to the native ligand. The pose with the lowest RMSD was not selected by the procedure explained in section 5.3 because it has a very low SC value (Table 5.3). That structure is showed in Figure 5.14 only for didactic purposes, and it was labeled as ‘best’.

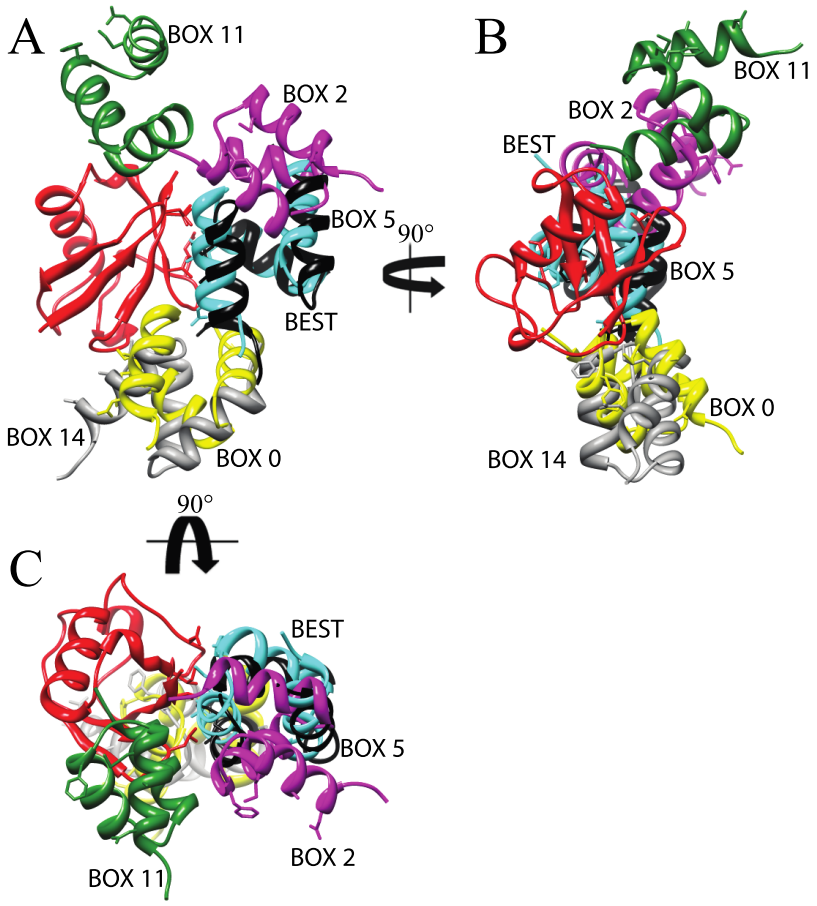

Figure 5.14: Selected representative poses for the Ubiquitin / Ubiquitin ligase complex.

'Best', according to ZDOCK, is the pose closest to the native-like, but it is not the native-like, it is only a regular complex; it is ~3 Å away from the native structure and it also has a very poor shape complementarity of 0.45; from the results obtained from the Barnase / Barstar complex and RNase / Barstar complex we know that the native-like pose has the highest SC value after the native and the RMSD value is ~2 Å. ZDOCK, FTDOCK (Gabb, Jackson et al. 1997), and Cluspro (Comeau, Gatchell et al. 2004; Comeau, Gatchell et al. 2004) did not find another pose closer to the native with an RMSD value lower than 3 Å; only

Rosetta Dock (Gray, Moughon et al. 2003; Lyskov and Gray 2008), found 89 poses very close to the native structure with an RMSD value below 3 Å.

For each pose, the respective values of BSA, SC and RMSD with respect to the native ligand are shown in Table 5.3.

Table 5.3: Ubiquitin / E3 Ubiquitin ligase representative poses: List of the chosen poses of the Ubiquitin / E3 Ubiquitin ligase complex with their respective BSA, SC and RMSD value versus the ligand native conformation.

	BSA (Å^2)	SC	RMSD (Å)
NATIVE	808	0.81	0.0
BOX0	997	0.54	20.3
BOX2	608	0.56	18.0
BOX5*	757	0.55	4.8
BOX11	779	0.59	31.5
BOX14	773	0.54	25.5

The ELNEDIN approach was done using the rigid elastic network of R_C = 1.2 nm and k_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} (1.2/1000). All the steps of how the PMF were obtained are explained in sections 2.2.4 and 2.2.5.

The resulting free energy profile (Figure 5.15) for each pose shows that box5 or native-like pose has a very different free energy profile from the native structure. It has a difference of free energy of ~5 kcal/mol with respect to the native (Figure 5.15 C). Box 5 has a SC value similar to the other poses and it has the lowest BSA value among all the other chosen poses. The two-peak surface at its bottom indicates that likely there is a range of conformational isomers (Tsai, Kumar et al. 1999). This feature shows that the native-like pose cannot be

differentiated from the other representative poses. Box 0 and box 14 both show a SC value of 0.54, the lowest value between the chosen poses. Box 0 has the highest BSA of 997 $Å^2$. Box 0 (Figure 5.15 A) and box 14 (Figure 5.15 E) have similar free energy profiles as the one obtained by the native conformation (black). The two plots have a minimum in the same position as the free energy profile of the native conformation (black) and, also, the depth of the well is very similar among the two of them.

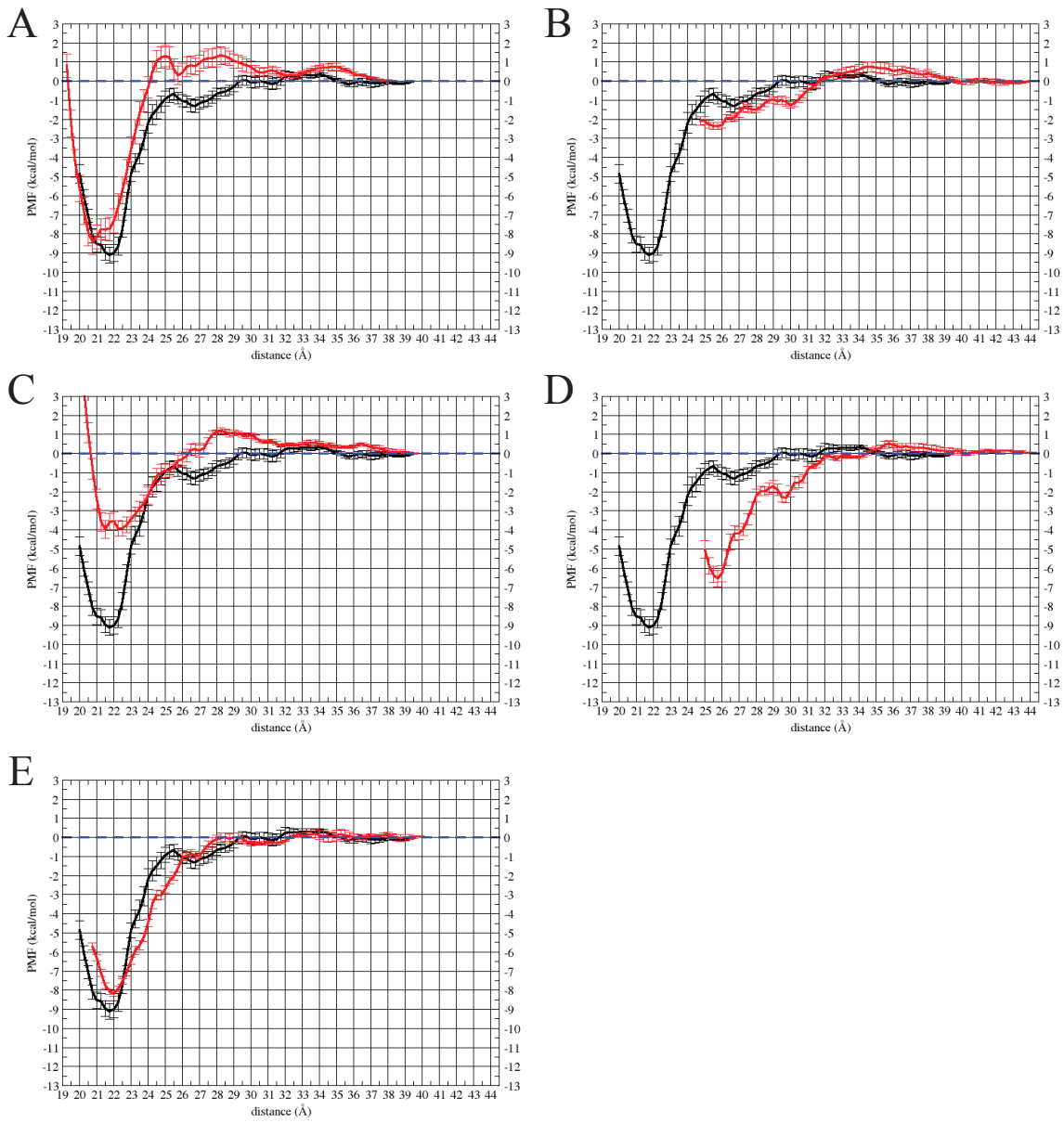

Figure 5.15: Free energy profile comparison of the dissociation of the Ubiquitin / Ubiquitin ligase chosen poses (red) versus the native conformation (black): (A) Box 0, (B) Box 2, (C) Box 5 or 'native-like', (D) Box 11, (E) Box 14.

To check wether the results are a force field issue, potential of mean force for box 0, box 14, box 5 and native were calculated using the atomistic approach (Figure 5.16). The resulting free energy profiles show that the there are force field

issues because Box 0 (Figure 5.16 A) is also picked as the best structure after the native conformation by the atomistic approach. Box 14 (Figure 5.16 C) was successfully differentiated by the atomistic approach indicating that this might be a drawback of the ELNEDIN approach. The free energy of binding of box 5 (Figure 5.16 B) is similar to the one obtained by the ELNEDIN approach.

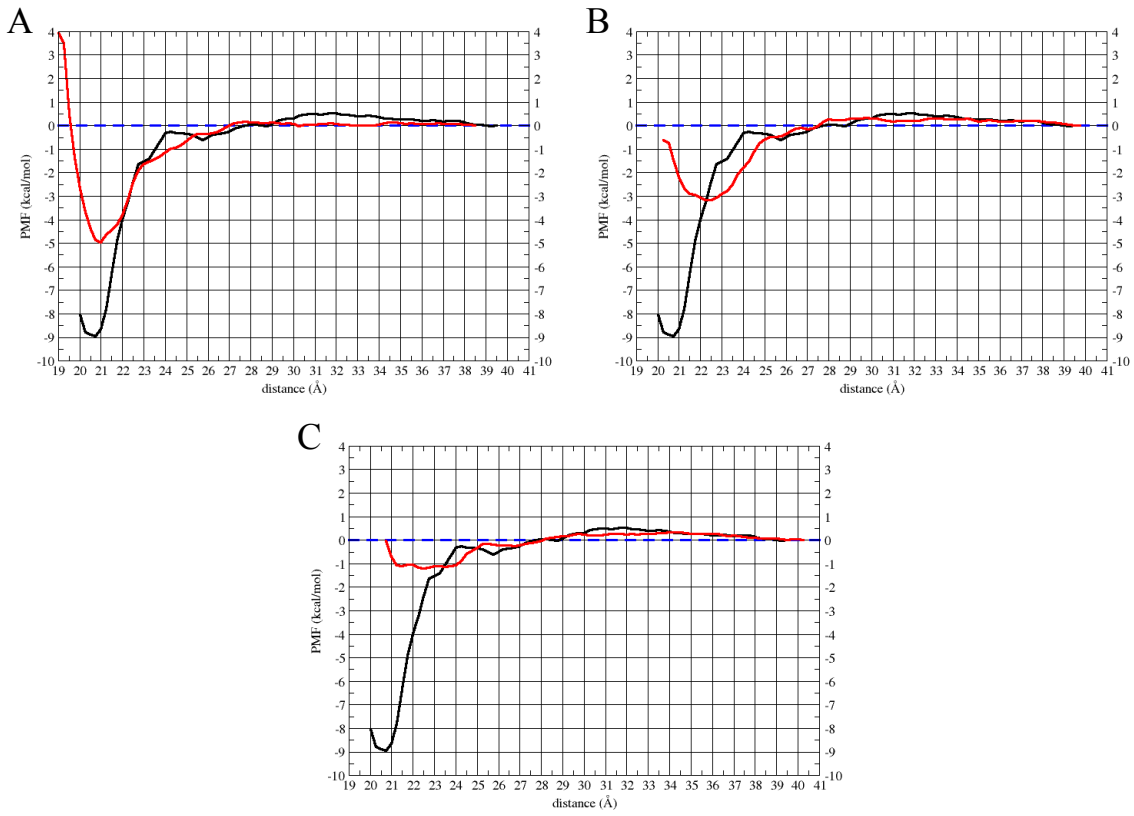

Figure 5.16: Comparison of the dissociation of the ubiquitin / ubiquitin ligase chosen poses versus (red) versus the native conformation (black) obtained by atomistic approaches: (A) Box 0, (B) Box 5 or 'native-like', (C) Box 14, (D) 'Best'.

The literature says that 2OOB is the only ubiquitin / ubiquitin ligase complex (Figure 5.17 A) that binds in a way different than the other ubiquitin / ubiquitin ligase complexes (i.e. PDB I.D.: 1WR1) (Kozlov, Nguyen et al. 2007;

Kozlov, Peschard et al. 2007; Peschard, Kozlov et al. 2007) (Figure 5.17 D). In the two cases the interactions in the interface are hydrophobic but 2OOB is interacting through helix 1 while 1WR1 is interacting with the loop of helix 1 and helix 2, and helix 3, making a hydrophobic patch. Figure 5.17 B shows the comparison between the representative structure of box 0 (gold) and the representative structure of box 14 (gray). They look similar, the reason why these two structures end up in similar boxes is because the position of the center of mass of these structure that in one case fell inside box 0, very close to box 14; and in the other case fell inside box 14, very close to box 0. They may appear similar but the interacting atoms are different. Figure 5.17 C might suggest two possibilities: 1) Box 0 may be interacting in the same way as 1WR1 or 2) Box 0 is taking the orientation of 1WR1 sometime along the simulation.

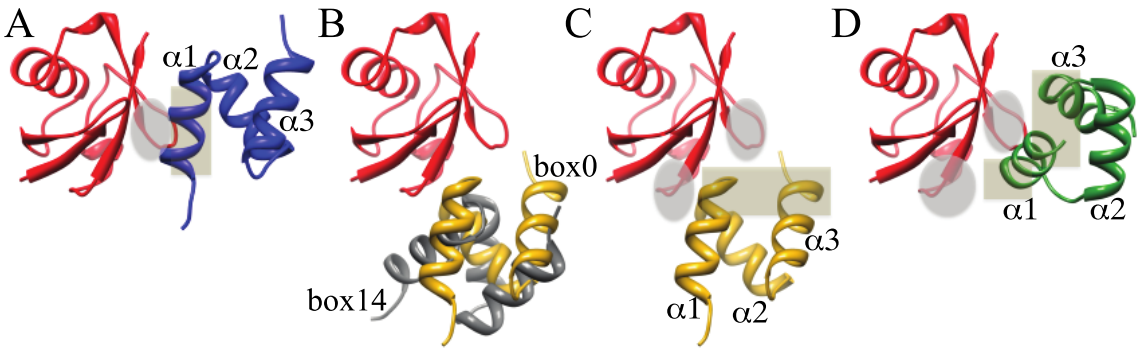

Figure 5.17: Comparison of the interactions of the different Ubiquitin / Ubiquitin ligase complexes: Cartoon representation of A) 2OOB native. B) Comparison of box0 (gold) and box14 (gray). C) 2OOB box 0. D) 1WR1. The hydrophobic hotspots of Ubiquitin (red) are encircled in light gray. The hydrophobic patches of E3 Ubiquitin ligase are enclosed in khaki rectangles.

The RMSD evolution of each trajectory discards the second hypothesis; the RMSD is mainly flat and below 2 Å in the first 3 Å of separation.

We have defined an 'interaction' as the atoms that are separated by less or equal to the sum of their respective van der Waals radii (Figure 5.18). In the two cases, box 0 and 1WR1, the interactions are hydrophobic:

- There are charge residues, but they are pointing away from the interface.
- There are two defined regions of the ligand that are interacting: helix 1 and helix 2 loop and helix 3.
- In the first region, in 1WR1, it is possible to see that H68 (receptor) is making contact with the backbone of M103 (helix 1 of ligand) and the backbone of D102 (helix 1 of ligand).
- The equivalent in box 0 are the interactions among H112 (residues), the side chain of M12 (helix 1 of ligand), and the backbone of E14 (loop helix 1 and helix 2 of ligand). M12 is also interacting with G13 and G15.
- The receptor of 1WR1 has also a G47 that is interacting with the side chain of M103 (helix 1 of ligand) and the backbone of F105 (helix 1 and helix 2 loop of the ligand).
- The equivalent of G47 in 2OOB box 0 is G91 but it is interacting with A43 and F44 in helix 3.
- The equivalent of F105 is Y16, but, in 2OOB box 0, the backbone of Y16 is interacting with H112.
- The Cγ of the ligand residue Q123 in helix 3 is interacting with V70 and L8.

- 2OOB box 0 has an additional hydrophobic contribution coming from the interaction of F89 and A90 with G15 and A17 (loop of helix 1 and helix 2).

Because of the similar regions with equivalent residues between 2OOB box 0 and 1WR1, we can say that the protein members of box 0 are interacting the same way as the other native ubiquitin complexes from other species.

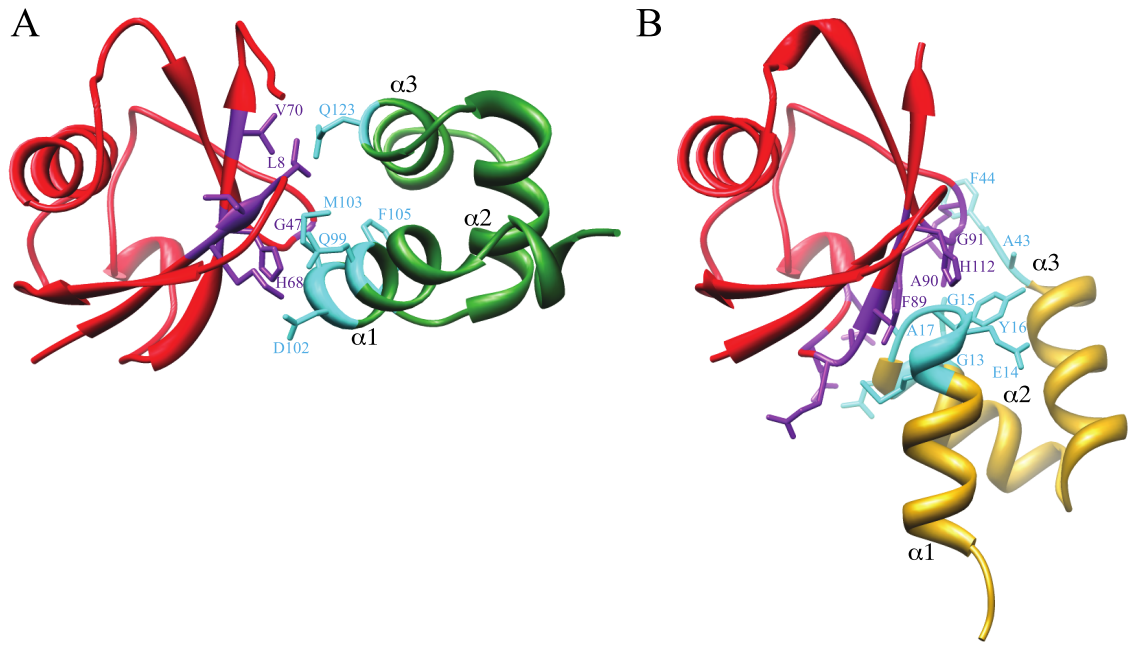

Figure 5.18: Interactions between ubiquitin and ubiquitin ligase in 1WR1 and 2OOB box 0: Cartoon representation of A) 1WR1. B) 2OOB box0.

The main difference between the Ubiquitin / Ubiquitin ligase complex with respect to the Barnase /Barstar complex and RNase / Barstar complex is that the interface of the Ubiquitin / Ubiquitin ligase is mainly hydrophobic; while the interface of the other two complexes are polar and charged. The literature says

that in the Barnase / Barstar complex and also, weakly, in the RNase / Barstar the electrostatics play an important role. This observation suggests that there might be something in the way of how the ELNEDIN approach is treating the hydrophobic interactions that leads to them becoming very attractive and favorable. In order to disrupt this attraction, we decided to switch from regular CG water to small CG water. Previously we saw how the use of small CG water destabilizes the free energy profile (section 4.3.2).

Figure 5.19 shows the free energy profiles for the native conformation (black), box 0 (Figure 5.19 A, red), box 5 (Figure 5.19 B, red) and box 14 (Figure 5.19 C, red). It is possible to observe that the difference of the free energy value at the minimum of well is ~2 kcal/mol between native and box 0 (Figure 5.19 A). 2 kcal/mol is a significant difference in comparison with the standard error displayed in Figure 5.19. The difference of the free energy value at the minimum of well is ~4 kcal/mol between native and box 14 (Figure 5.19 B). The difference of the free energy of binding between box 0 and box 14 is ~2.5 kcal / mol. The results, obtained for the free energy profile of box 0 and box 14 using the ELNEDIN approach with small water, are in agreement with the free energy profiles obtained by using the atomistic approach. The free energy profile obtained for box 5 differs from what was obtained with the atomistic approach. The ELNEDIN approach using small water recognized box 5 as the native-like.

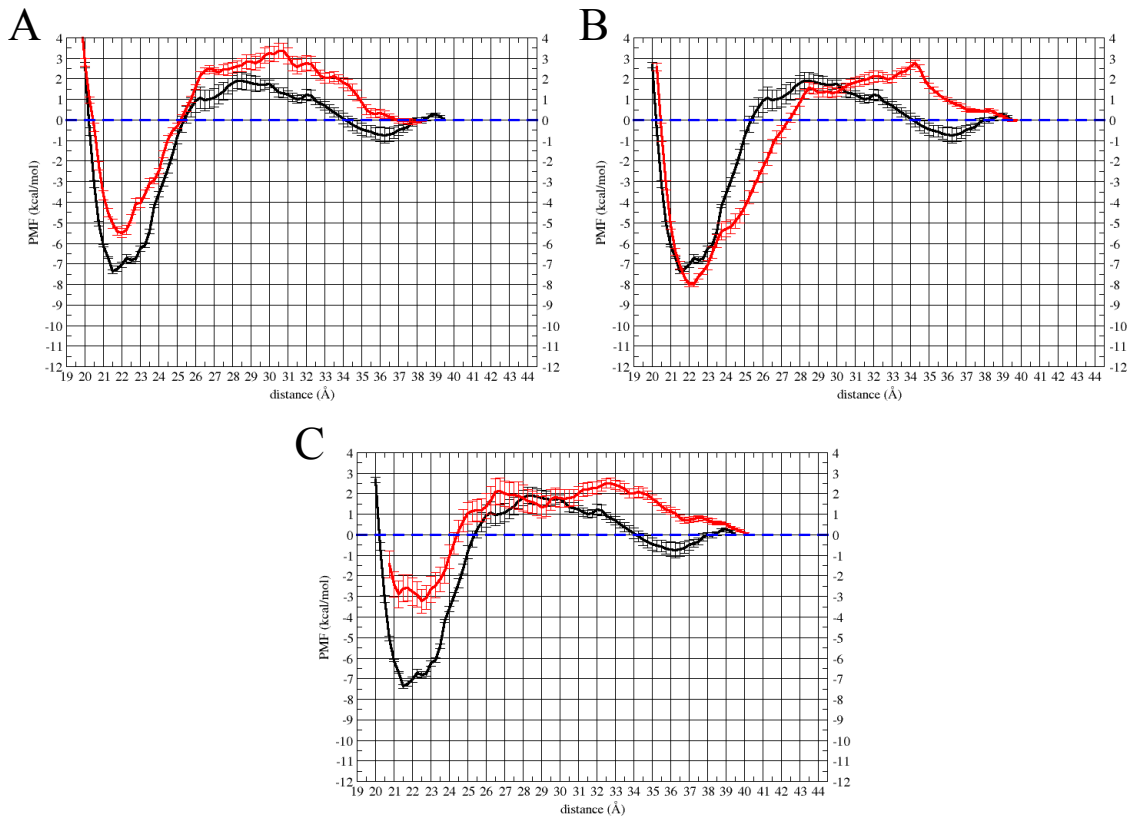

Figure 5.19: The effect of the small water: A) Box 0, B) Box 5, C) Box 14.

Figure 5.20 shows the average RMSD per bin for box 14 with respect to the initial structure of box 0 for the first 3 Å. One possible explanation of why the free energy profile of box 14 is not differentiated from the one obtained for box 0 is: when the ELNEDIN regular approach with regular water is used (black), the conformation of box 14 gets closer to the conformation of box 0 in the first 2 Å. When the atomistic approach (green) or the ELNEDIN approach with small water (red) is used the conformation of box 14 does not come closer to the conformation of box 0.

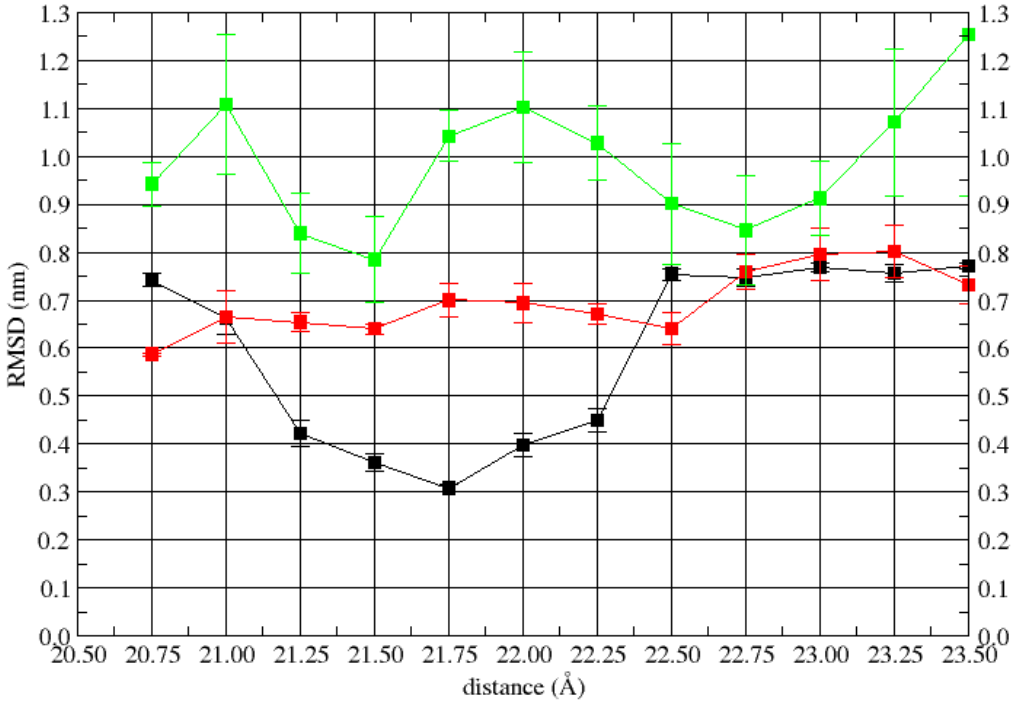

Figure 5.20: Average RMSD per bin for box 14 with respect to box 0: The ELNEDIN approach with regular water is depicted in black, with small water is in red and the atomistic approach in green.

Looking for an explanation for why the ELNEDIN approach using small water recognized box 5 as the native-like structure, we have counted the number of water molecules in the interface. The difference of the average of the numbers of small waters minus the number of regular waters was plotted per bin for the first 3 Å (Figure 5.21). But the plot did not show any useful information. In almost all the cases box 0 has the biggest difference. Box 14 the smallest; and at least for the first 2.5 Å the difference of interface waters for native and box 5 is not significant.

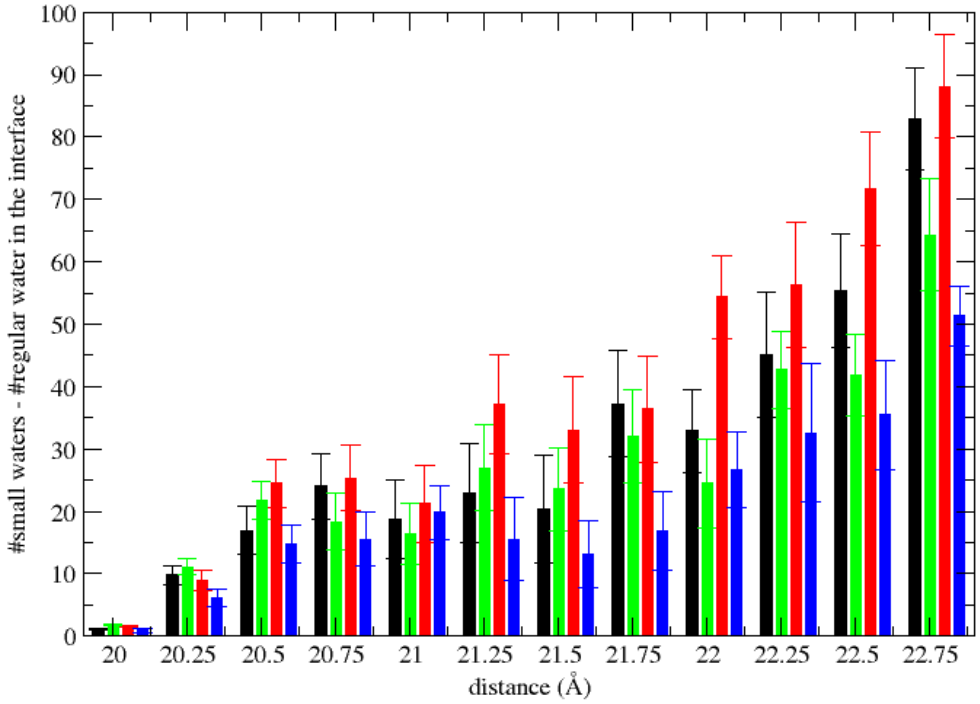

Figure 5. 21: Difference of the interface waters for each complex: Native (black), Box 5 (green), Box 0 (red) and Box 14 (blue).

Figure 5.22 shows how the average SC value per bin behaves for the first 3 Å. The SC values for the four complexes have the same tendency and it shows that the use of a small water (red) affects negatively the shape complementarity of the interface.

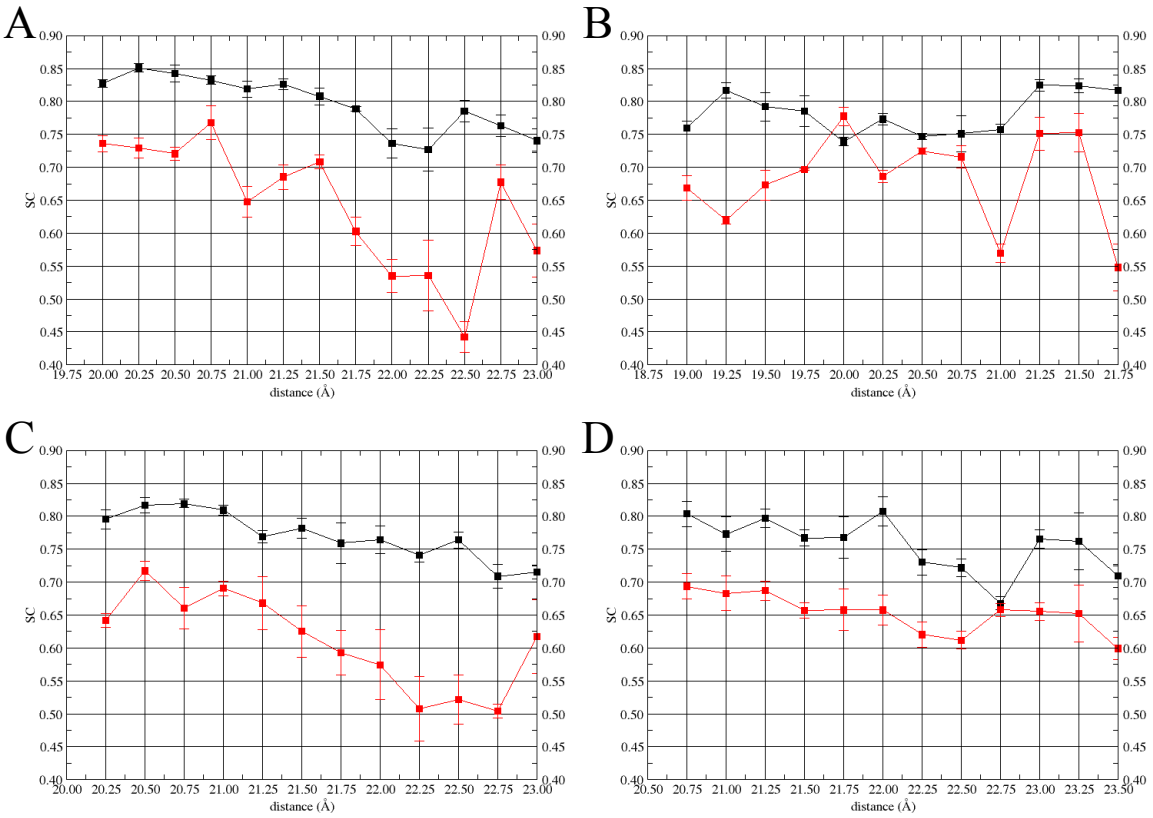

Figure 5.22: Average SC per bin: regular water (black) and small water (red). A) Native, B) Box 0, C) Box 5, D) Box 14.

Finally we checked the average RMSD value per bin for the first 3 Å for box 5 with respect to the native conformation (Figure 5.23). The conformation of box 5 leans slightly towards the native but it does not seem to be enough to account for box 5, which was picked over the other complexes as the native-like by the ELNEDIN approach using small water.

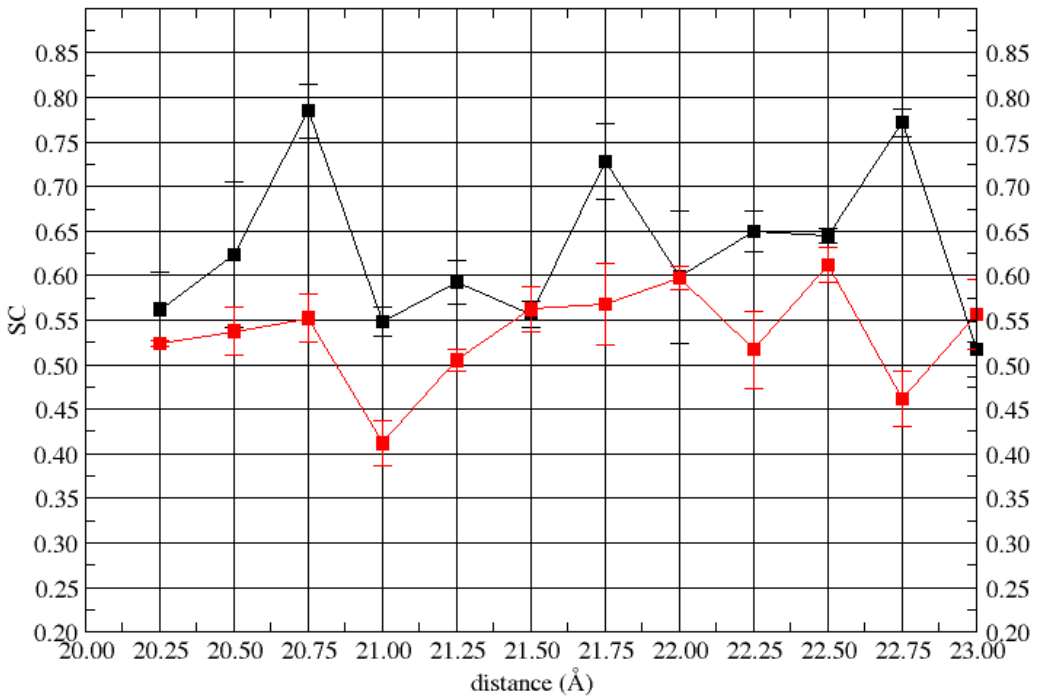

Figure 5.23: Average RMSD per bin for box 5 with respect to the native conformation: The ELNEDIN approach with regular water is depicted in black and with small water is in red.

The reason why box 5 was picked over the other complexes, when a small water is used, is still not clear.

5.4.4 The Bound – Bound form of the Nuclease A (NucA) / Nuclease inhibitor A (NuiA) complex case:

The NucA / NuiA complex is another example of a complex with an interface composed of polar and charged residues. This case was performed with the ultimate goal of obtaining preliminary results for future work. After performing the steps explained in section 5.4 (a detailed description of each step is provided in appendix A.4), the following poses were chosen and named after the location of the box in which they were found: box 0, box 5, box 13 (this pose corresponds to the native-like pose), box 16 and box 22 (Figure 5.24).

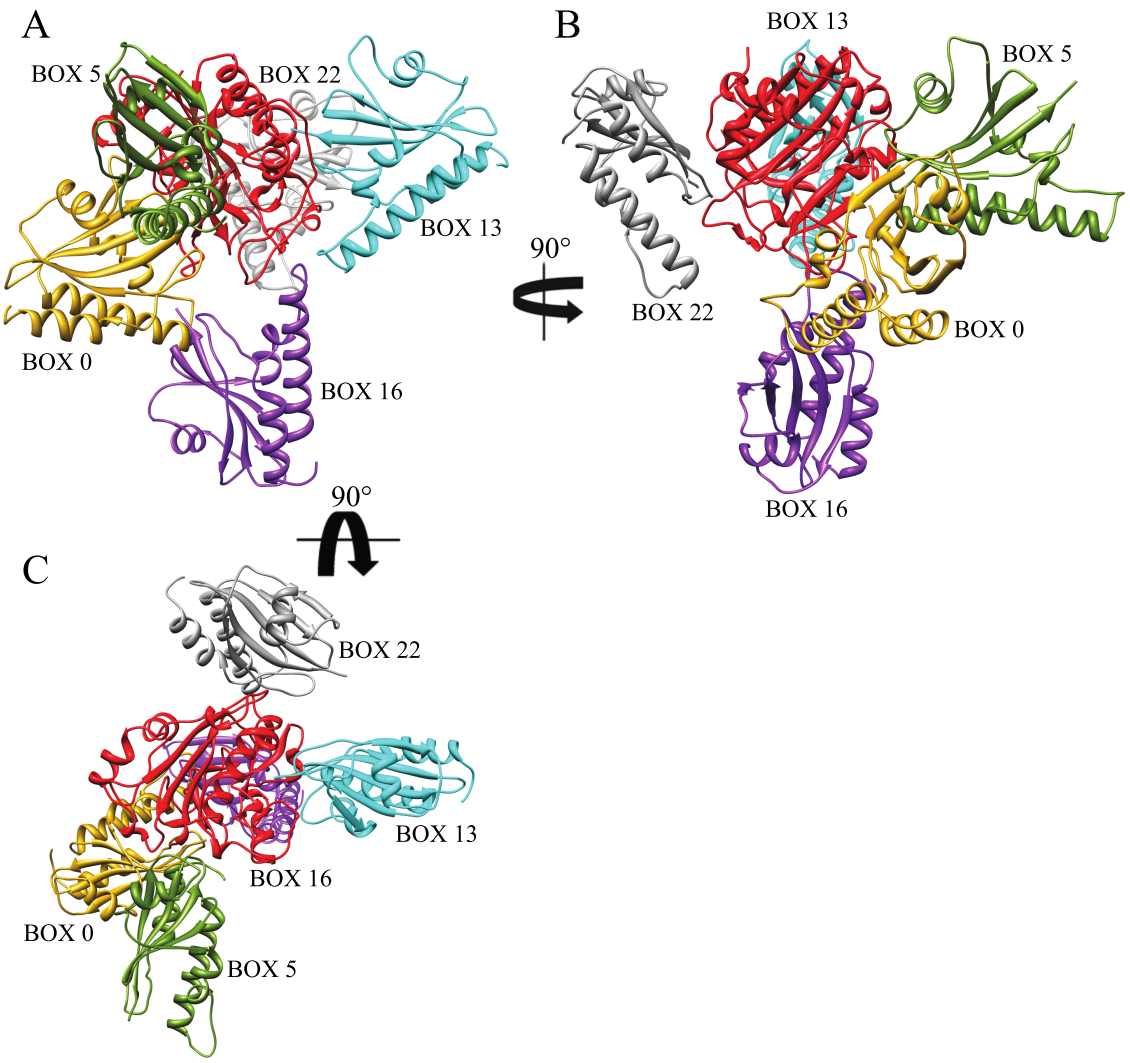

Figure 5.24: Selected representative poses for the NucA / NuiA complex.

For each pose, the respective values of BSA, SC and RMSD with respect to the native ligand are shown in Table 5.4.

Table 5.4: The Bound-Bound NucA / NuiA complex representative poses: List of the chosen poses of the bound-bound form of the NucA - NuiA complex with their respective BSA, SC and RMSD value versus the ligand native conformation.

	BSA (Å^2)	SC	RMSD (Å)
NATIVE	1668	0.64	0
BOX0	1631	0.48	64.2
BOX5	1182	0.53	55.3
BOX13*	1889	0.61	1.1
BOX16	852	0.62	59.5
BOX22	1292	0.63	45.0

The ELNEDIN approach was done using the rigid elastic network of R_C = 1.2 nm and k_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} (1.2/1000). All the steps of how the PMF were obtained are explained in sections 2.2.4 and 2.2.5. The SC value of box 13 (SC = 0.61) or native-like is very close to the native one (SC = 0.64). It is not the second higher because box 16 (SC = 0.62) and box 17 (SC = 0.63) have also similar SC values, but the BSA of box 13 is the highest and the closest to the native (BSA = 1889). The native-like pose (Figure 5.25-C) has a free energy profile very similar to the free energy profile of the native conformation, with a small shoulder ~ 4 Å away from the minimum of the well and a minimum free energy value ~ 3 kcal/mol of difference from the native conformation, but in the bottom it has two minima. Suggesting that there are two conformations: the bound-bound form of NucA – NuiA complex and its native-like conformation 1.1 Å closer. Tsai et al (Tsai, Kumar et al. 1999) suggests that flexible proteins have the tendency to have rugged funnel bottoms corresponding to their conformational isomers. This result indicates that the ELNEDIN approach was able to identify the native-like conformation.

The free energy profile for the other boxes shows different shapes and lower binding free energy with respect to the free energy profile of the native or native-like pose.

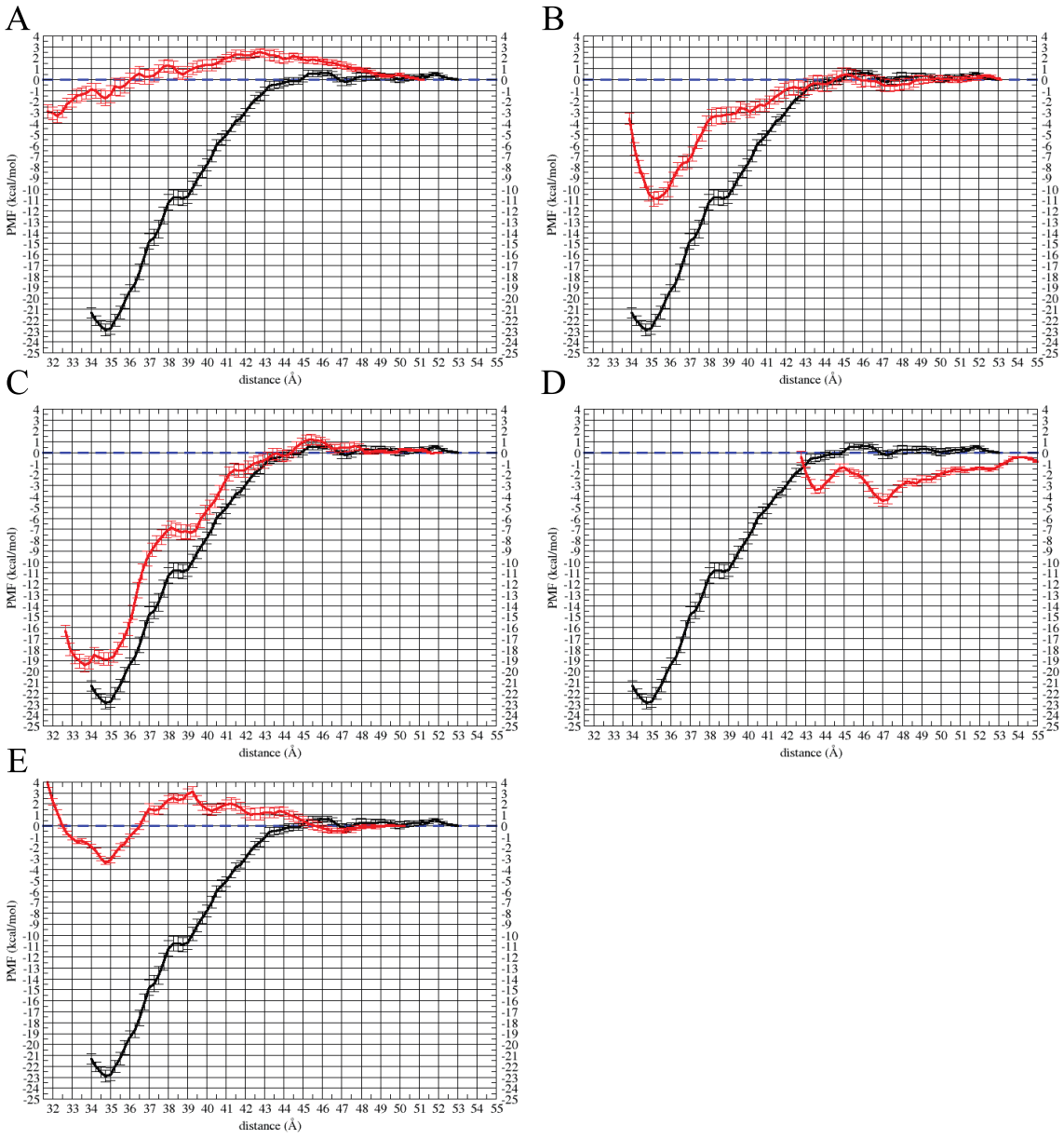

Figure 5.25: Free energy profile comparison of the dissociation of the bound-bound form of NucA / NuiA complex representative poses versus the native conformation: A) Box 0, B) Box 5, C) Box 13 or 'native-like', D) Box 16, E) Box 22.

Of the four cases studied, in three of them, the native-like pose was successfully identified by ELNEDIN, and that is a good result. The only failed case is one with a hydrophobic interface.

5.5 Can the unbound form of the ligand bind to the receptor specifically?

To test if the unbound form of the ligand can bind to the receptor specifically, the bound – unbound form of the NucA / NuiA complex was used. The RMSD value between the complexed and the free form of the receptor is 0.34 Å for all Cα atoms, and the RMSD value between the complexed and the free ligand is 3.5 Å. After performing the steps explained in section 5.4 (a detailed description of each step is given in appendix A.5), the following poses were chosen and named after the location of the box in which they were found: box 0, box 5, box11, box 13 (this pose corresponds to the native-like pose), box 16 and box 22 (Figure 5.26).

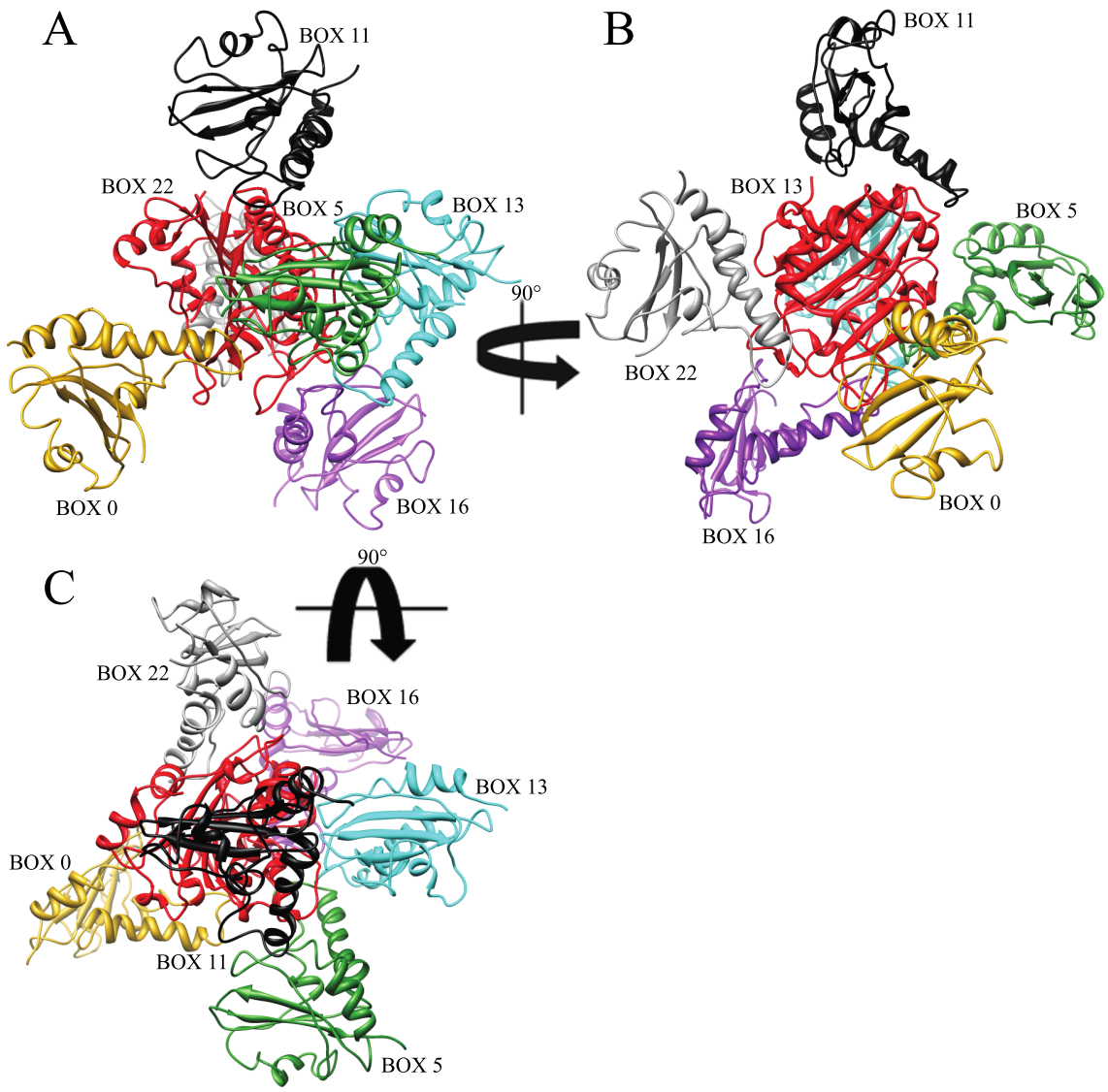

Figure 5.26: Selected representative poses for the NucA / NuiA complex.

For each pose, the respective values of BSA, SC and RMSD with respect to the native ligand, are shown in Table 5.5.

Table 5.5: The Bound-Unbound NucA / NuiA complex representative poses: List of the chosen poses of the bound-unbound form of the NucA - NuiA complex with their respective BSA, SC and RMSD value versus the ligand native conformation.

	BSA (Å^2)	SC	RMSD (Å)
'NATIVE'	1009	0.22	0
BOX0	1168	0.53	66.9
BOX5	1465	0.47	39.8
BOX11	1340	0.50	45.3
BOX13*	1341	0.57	4.8
BOX16	1167	0.49	44.6
BOX22	1892	0.46	51.8

The ELNEDIN approach was done using the rigid elastic network of R_C = 1.2 nm and k_{SPRING} = 1000 kJ.mol^{-1}.nm^{-2} (1.2/1000). All the steps of how the PMF were obtained are explained in sections 2.2.4 and 2.2.5. The 'native' structure, because it is not a real structure in nature, possesses a very low SC value of 0.22. The SC value for box 13 or 'native-like' is the highest value (SC = 0.57). The native-like pose (Figure 5.27 D) has a free energy profile very similar to the free energy profile of the native conformation, with a minimum free energy value ~ 3 kcal/mol of difference from the native conformation (black). The free energy profile for the other boxes shows different shapes and lower binding free energy with respect to the free energy profile of the native or native-like pose

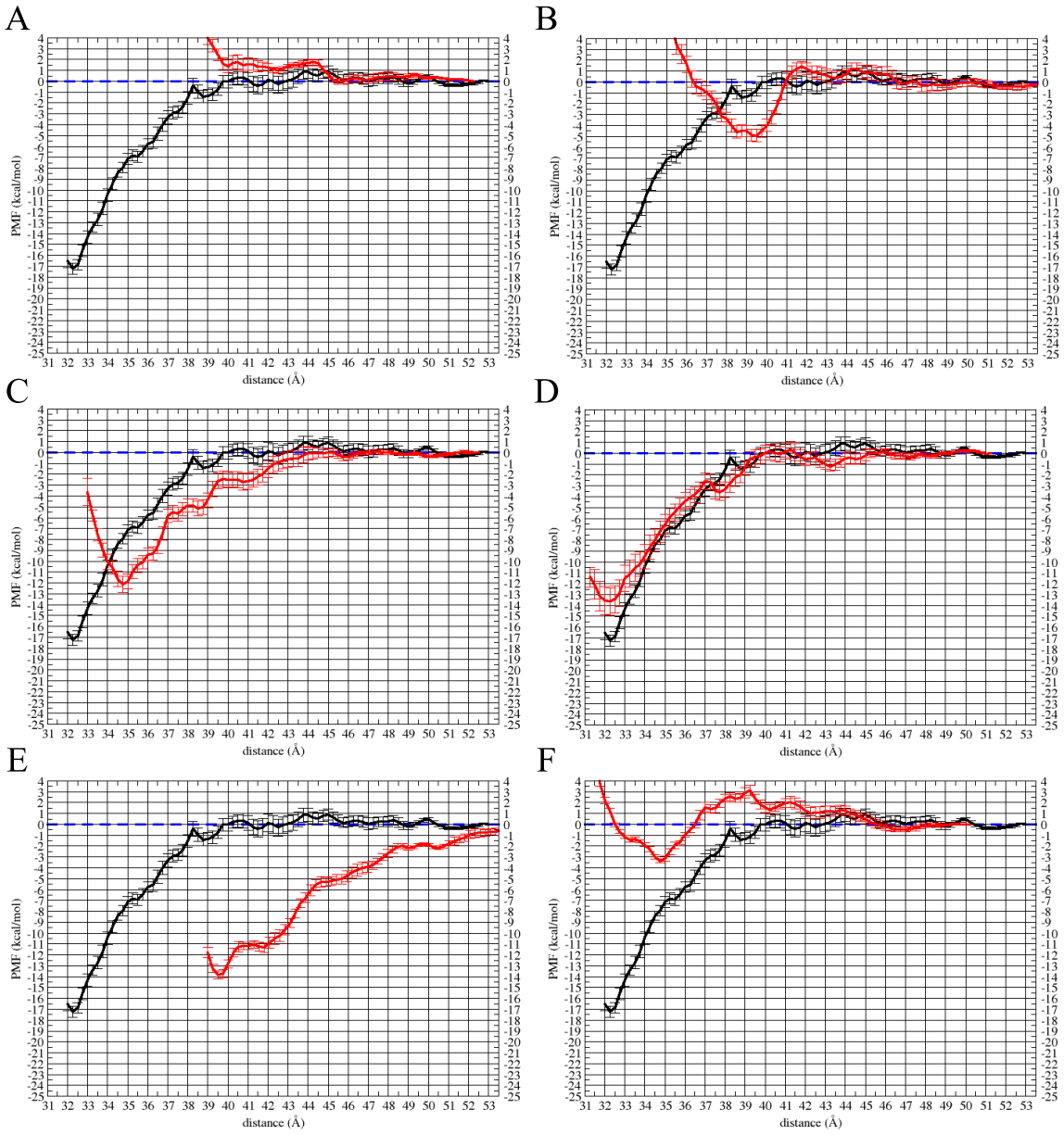

Figure 5.27: PMF comparison of the dissociation of the bound - unbound form of NucA / NuiA complex representative poses versus the native conformation: A) Box 0, B) Box 5, C) Box11, D) Box 13 or 'native-like', E) Box 16, F) Box 22.

This result shows that the ELNEDIN approach was able to identify the native-like conformation. Thus indicating, at least for one case, that the unbound form of the ligand is able to bind to the receptor specifically.

6

Conclusions

In this work we have tested the ability of the ELNEDIN approach to identify non-native interfaces and recognize only the native or native-like pose.

We started by validating and testing the ELNEDIN approach by comparing the free energy profiles obtained by the ELNEDIN approach against the atomistic models. It was found that it is possible to obtain accurate energy-profiles using the ELNEDIN approach.

We obtained insight on the shape of the free energy landscape around a protein receptor. It is possible to see the gradient, or funnel-like shape, where the bottom of the funnel is the global and lowest minimum and to observe that it increases smoothly. The ligand might be able to rotate -60° on its center of mass around the x-axis, and 60° or -60° around the y-axis without affecting much the overall shape of the free energy map.

By modifying the Lennard-Jones interactions of the Martini 2.1 force field, and through the replacement of the regular CG water for a small CG water, we were able to explore the importance of the solvent in the shape of the free energy profile.

Finally, we demonstrated that in 3 out of 4 cases the ELNEDIN approach was able to recognize the native interface. The three successful cases have hydrophilic interfaces. This is very promising for a future application of the ELNEDIN approach in the quick recognition of native hydrophilic interfaces. The one case left corresponds to the only case, the Ubiquitin / Ubiquitin Ligase case, where the interface between the two proteins is hydrophobic; the

ELNEDIN approach likely failed due to force field issues related to the treatment of hydrophobic residues. One of the successful cases corresponds to a pair of proteins that undergo large conformational changes upon complexation, the NucA / NuiA complex; this successful result is a promising one for possible future work regarding the use of the ELNEDIN approach in allowing conformational changes in the complexation of two unbound proteins.

APPENDIX

A.1.- The Clustering of the Barnase / Barstar complex

The following distribution was obtained:

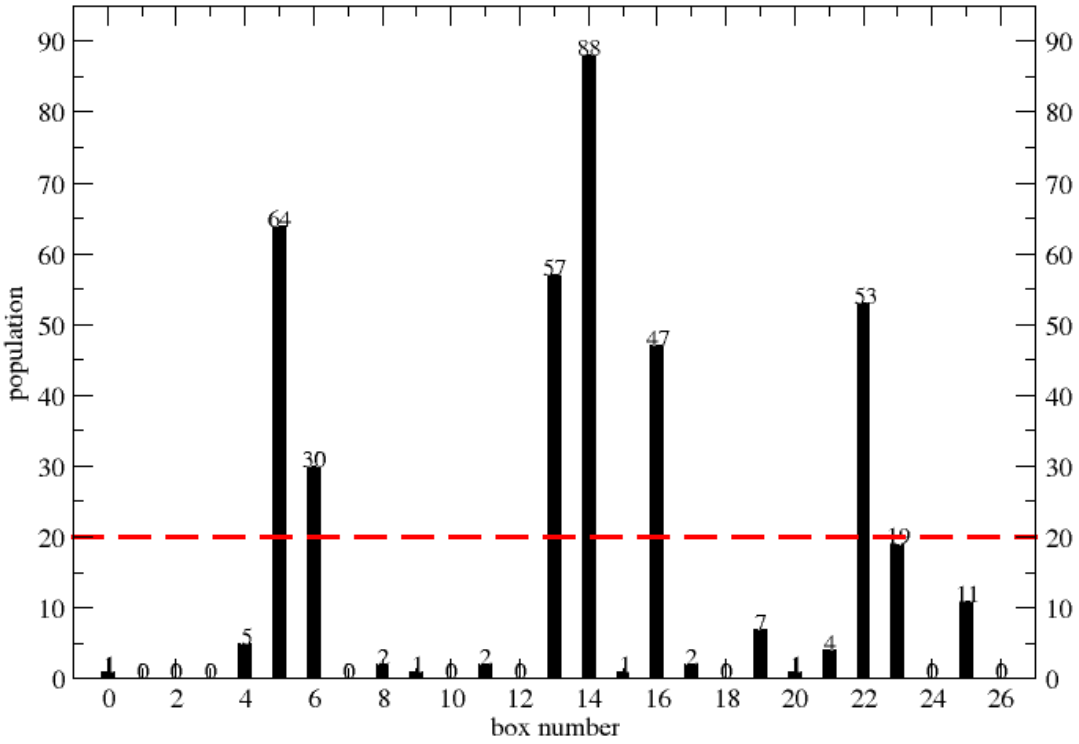

Figure A.1: Distribution of the Barnase / Barstar poses: The cut-off of 20 structures is marked by a red discontinuous line.

Using a cut-off of 20 structures:

Table A. 1: Distribution of the Barnase / Barstar poses

BOX NUMBER	# of structures	Average RMSD (nm)
5	64	2.24 +/- 0.87
6	30	1.93 +/- 0.80
13	57	1.81 +/- 0.49
14*	88	1.79 +/- 0.60
16	47	1.52 +/- 0.44
22	53	2.01 +/- 0.81

To each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. We have chosen the cluster with the largest number of members within a fixed cluster distance less than 10 Å. In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. As the procedure is straightforward for each box, only the data for one random box will be shown:

BOX 6:

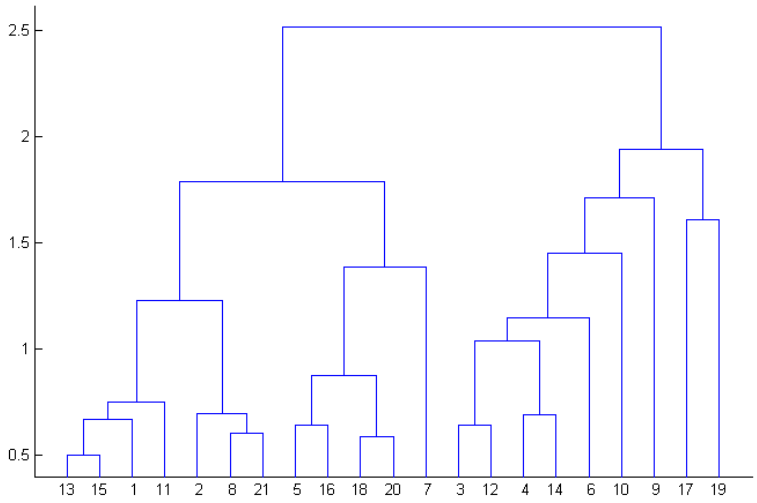

Figure A.2: Dendrogram obtained from the hierarchical clustering of poses inside box 6.

Table A.2: Hierarchical clustering of the poses inside box 6

Cluster	# of structures	Average rmsd (nm)
13_15_1_11	7	0.28 +/- 0.11
2_8_21	5	0.25 +/- 0.08
5_16_18_20	5	0.34 +/- 0.13
7	2	0.17 +/- 0.00
3_12	2	0.37 +/- 0.00
4_14	3	0.21 +/- 0.04
6	2	0.16 +/- 0.00
10	1	0
9	1	0
17	1	0
19	1	0

A.2.- The clustering of the RNase / Barstar complex

The following distribution was obtained:

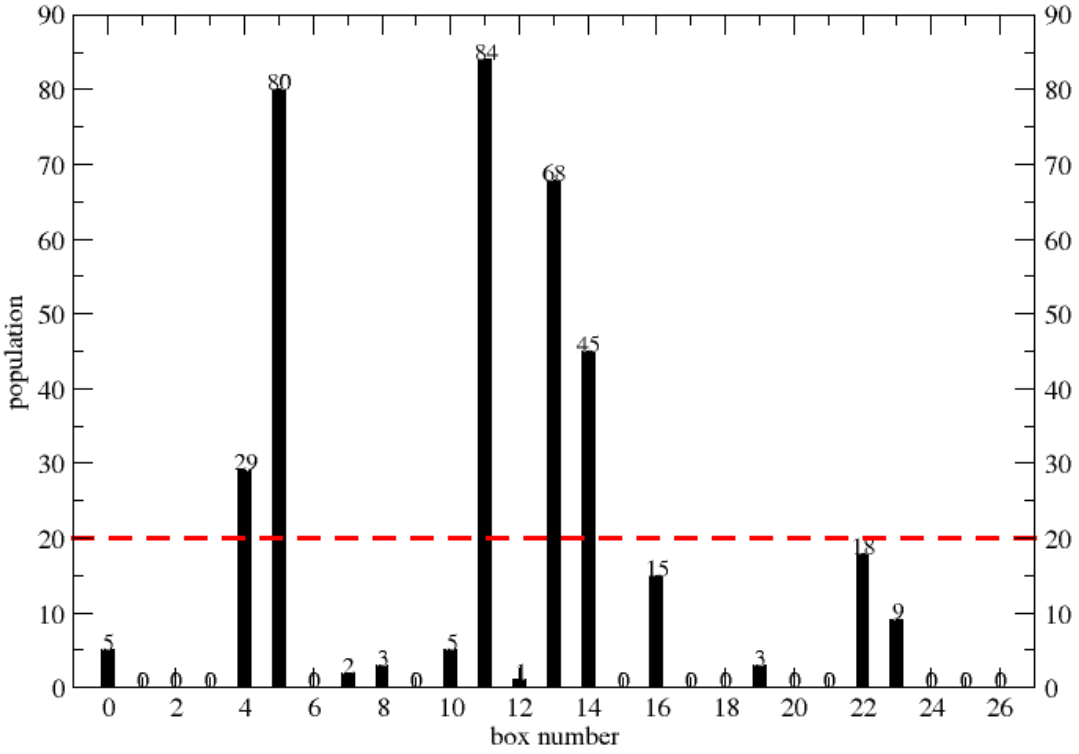

Figure A.3: Distribution of the RNase / Barstar poses: The cut-off of 20 structures is marked by a red discontinuous line.

Using a cut-off of 20 structures:

Table A.3: Distribution of the RNase / Barstar poses

BOX NUMBER	# of structures	Average RMSD (nm)
4	29	1.66 +/- 0.39
5*	80	2.06 +/- 0.6
11	84	1.99 +/- 0.6
13	68	2.00 +/- 0.55
14	45	1.69 +/- 0.55

To each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. We have chosen the cluster with the largest number of members within a fixed cluster distance less than 10 Å. In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. As the procedure is straightforward for each box, only the data for one random box will be shown:

BOX 14:

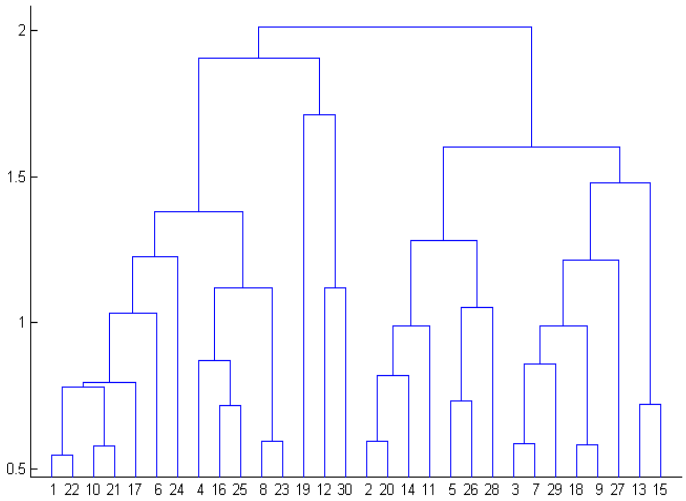

Figure A.4: Dendrogram obtained form the hierarchical clustering of the poses inside box 14.

Table A.4: Hierarchical clustering of the poses inside box 14

Cluster	# of structures	Average rmsd	Cluster	# of structures	Average rmsd
1_22_10_21_17	9	0.28 +/- 0.10	30	1	0
6	1	0	2_20_14_11	6	0.37 +/- 0.15
24	2	0.16 +/- 0.00	5_26	5	0.28 +/- 0.11
4_16_25	4	0.39 +/- 0.12	28	1	0
8_23	2	0.36 +/- 0.00	3_7_29_18_9	7	0.32 +/- 0.15
19	1	0	27	3	0.17 +/- 0.01
12	1	0	13_15	2	0.20 +/- 0.00

A.3.- The clustering of the Ubiquitin / Ubiquitin Ligase complex

The following distribution was obtained:

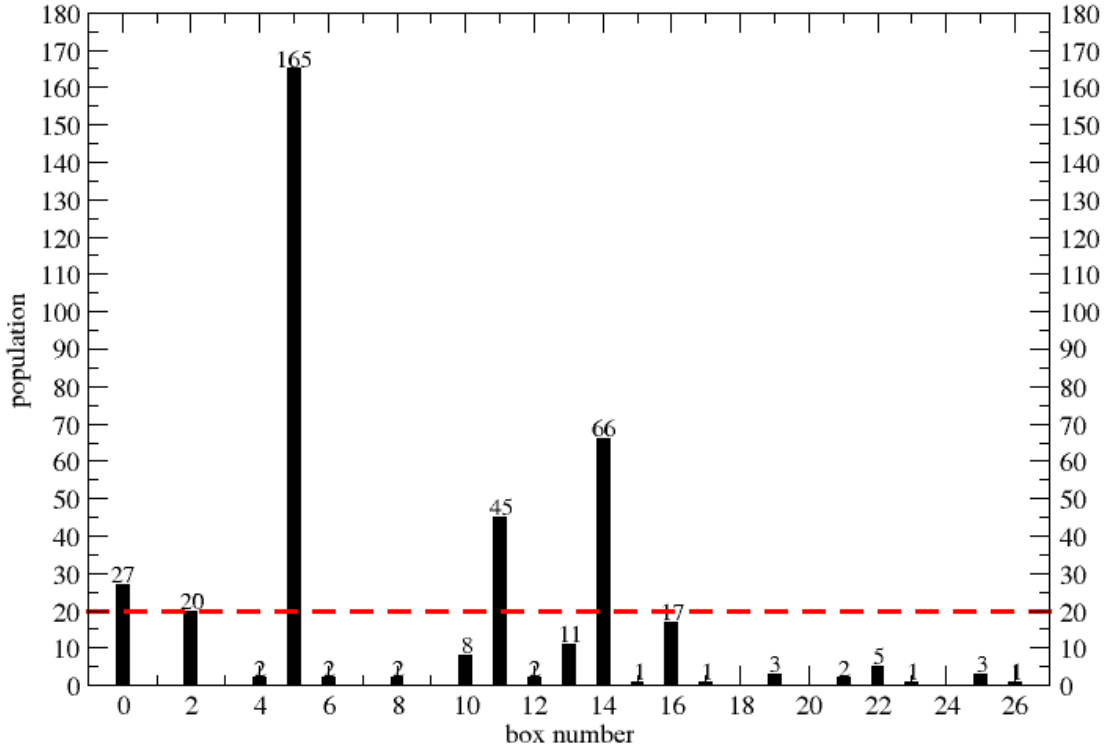

Figure A.5: Distribution of the Ubiquitin / Ubiquitin Ligase poses: The cut-off of 20 structures is marked by a red discontinuous line.

Using a cut-off of 20 structures:

Table A.5: Distribution of the Ubiquitin / Ubiquitin Ligase poses

BOX NUMBER	# of structures	Average RMSD (nm)
0	27	1.20 +/- 0.35
2	20	1.13 +/- 0.35
5*	165	1.78 +/- 0.54
11	45	1.79 +/- 0.65
14	66	1.77 +/- 0.57

To each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. We have chosen the cluster with the largest number of members within a fixed cluster distance less than 10 Å. In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. As the procedure is straightforward for each box, only the data for one random box will be shown:

BOX 0:

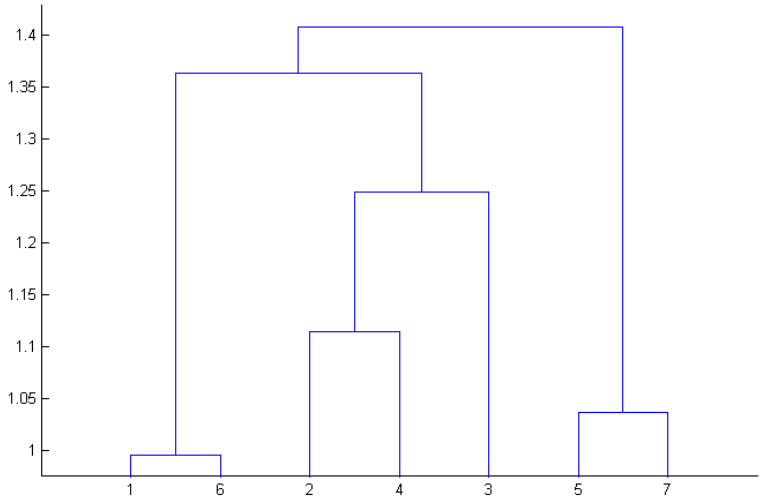

Figure A.6: Dendrogram obtained from the hierarchical clustering of the poses inside box 0.

Table A.6: Hierarchical clustering of the poses inside box 0

Cluster	# of structures	Average rmsd (nm)
1	4	0.16 +/- 0.04
6	5	0.33 +/- 0.13
2	9	0.33 +/- 0.13
4	3	0.16 +/- 0.05
3	1	0
5	4	0.27 +/- 0.15
7	1	0

A.4.- The clustering of the bound – bound form of the NucA / NuiA complex

The following distribution was obtained:

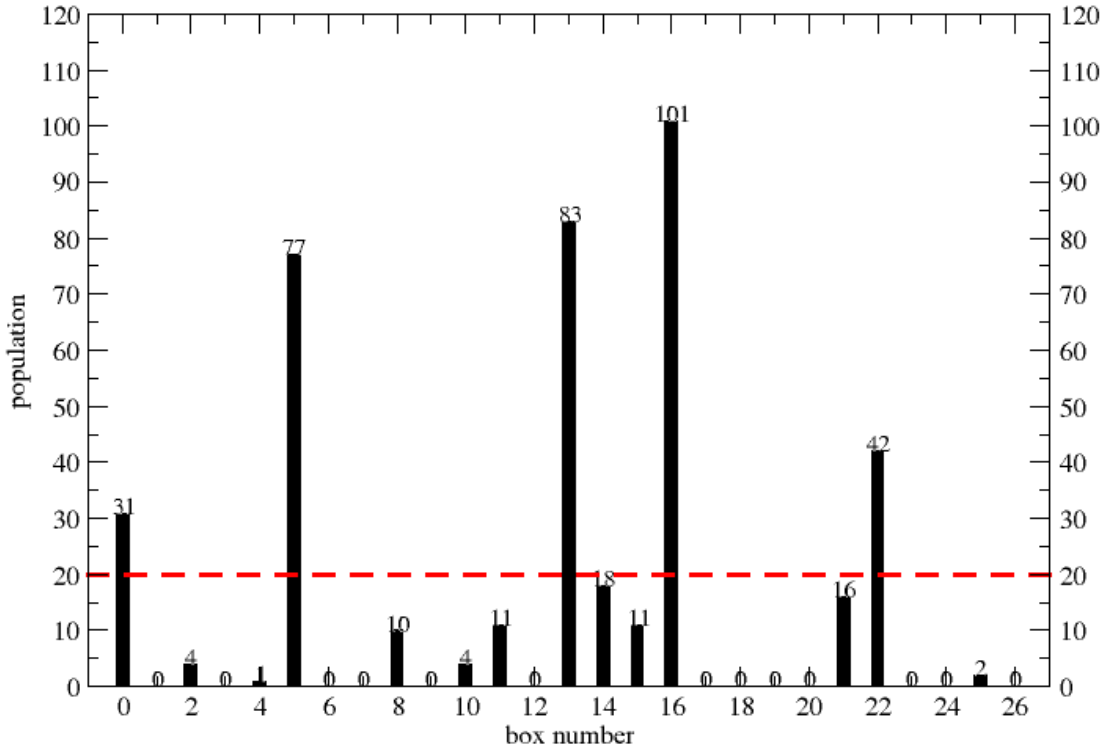

Figure A.7: Distribution of the bound – bound form of the NucA / NuiA poses: The cut-off of 20 structures is marked by a red discontinuous line.

Using a cut-off of 20 structures:

Table A.7: Distribution of the bound – bound form of the NucA / NuiA poses.

BOX NUMBER	# of structures	Average RMSD (nm)
0	31	2.86 +/- 0.86
5	77	2.75 +/- 0.70
13*	83	2.75 +/- 0.75
16	101	2.94 +/- 0.84
22	42	2.87 +/- 0.95

To each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. We have chosen the cluster with the largest number of members within a fixed cluster distance less than 10 Å. In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. As the procedure is straightforward for each box, only the data for one random box will be shown:

BOX 22:

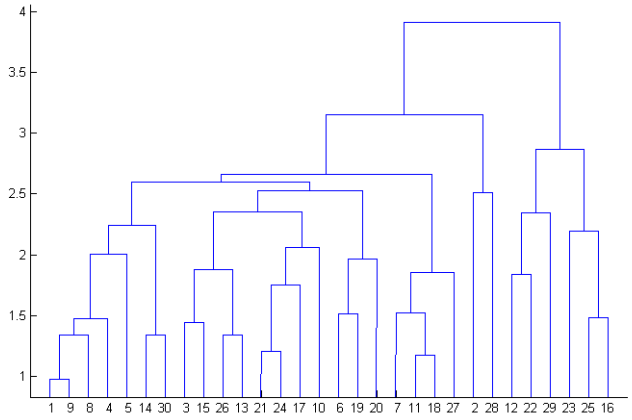

Figure A.8: Dendrogram obtained from the hierarchical clustering of the poses inside box 22.

Table A.8: Hierarchical clustering of the poses inside box 22

Cluster	# of structures	Average rmsd	Cluster	# of structures	Average rmsd
1	1	0	16	1	0
2	1	0	17	1	0
3	2	0.20 +/- 0.00	18	1	0
4	1	0	19	1	0
5	6	0.25 +/- 0.07	20	1	0
6	2	0.45 +/- 0.00	21	1	0
7	1	0	22	1	0
8	3	0.33 +/- 0.05	23	1	0
9	1	0	24	1	0
10	1	0	25	1	0
11	2	0.29 +/- 0.00	26	1	0
12	2	0.21 +/- 0.00	27	1	0
13	1	0	28	1	0
14	1	0	29	1	0
15	2	0.29 +/- 0.00	30	1	0

A.5.- The clustering of the bound – unbound form of the NucA / NuiA complex

The following distribution was obtained:

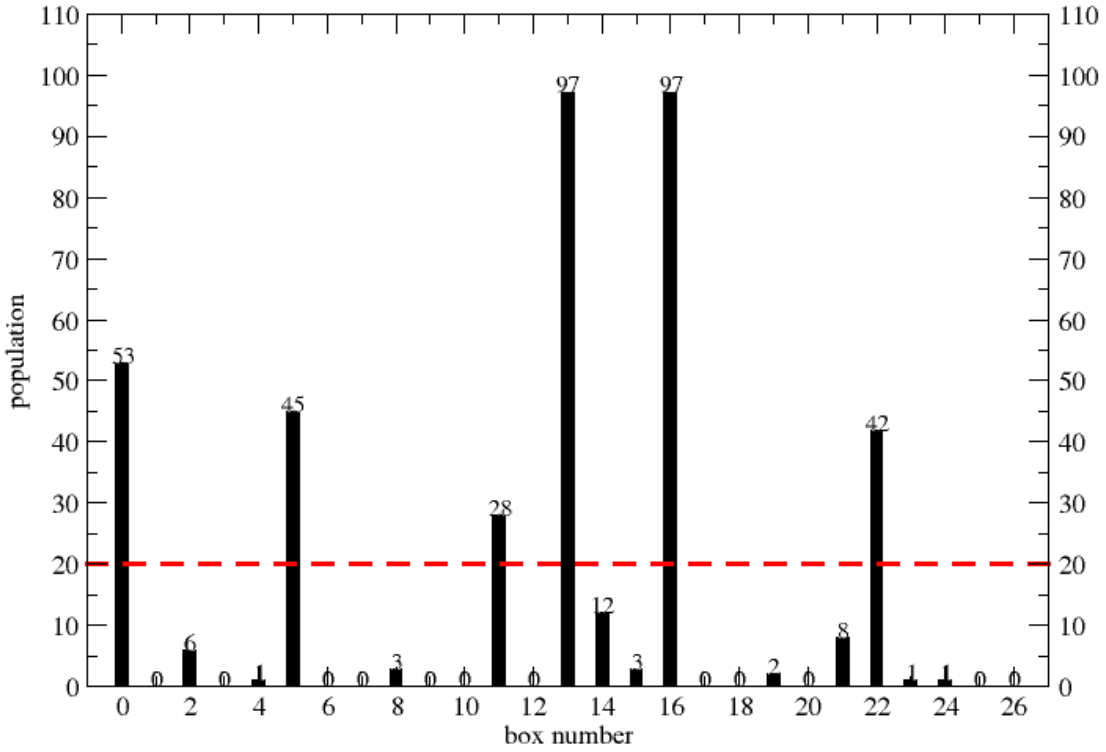

Figure A.9: Distribution of the bound – unbound form of the NucA / NuiA poses: The cut-off of 20 structures is marked by a red discontinuous line.

Using a cut-off of 20 structures:

Table A.9: Distribution of the bound – unbound form of the NucA / NuiA poses.

BOX NUMBER	# of structures	Average RMSD (nm)
0	53	3.49 +/- 1.41
5	45	2.93 +/- 0.97
11	28	2.69 +/- 1.00
13*	97	2.78 +/- 0.79
16	97	2.85 +/- 0.80
22	42	2.62 +/- 0.85

To each chosen box, an average linkage hierarchical clustering on the basis of pairwise RMSD was done. We have chosen the cluster with the largest number of members within a fixed cluster distance less than 10 Å. In the case that there were two or more clusters with the same number of members, we picked the one with the less average RMSD value among all members of that chosen cluster. As the procedure is straightforward for each box, only the data for one random box will be shown:

BOX 11:

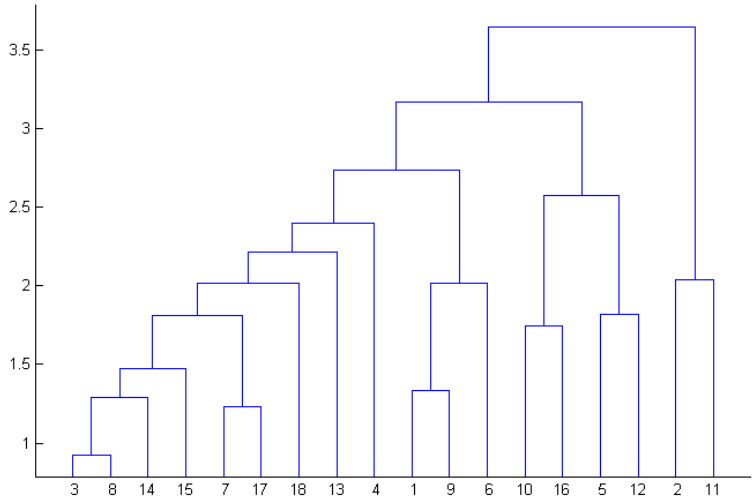

Figure A.10: Dendrogram obtained from the hierarchical clustering of the poses inside box 11.

Table A.10: Hierarchical clustering of the poses inside box 11

Cluster	# of structures	Average rmsd	Cluster	# of structures	Average rmsd
1	1	0	10	1	0
2	1	0	11	3	0.35 +/- 0.11
3+8	4	0.35 +/- 0.11	12	1	0
4	1	0	13	1	0
5	1	0	14	1	0
6	2	0.17 +/- 0.00	15	1	0
7	3	0.31 +/- 0.08	16	1	0
9	1	0	17	2	0.21 +/- 0.00
			18	1	0

REFERENCES

Alberts, B. (1998). "The cell as a collection of protein machines: preparing the next generation of molecular biologists." Cell **92**(3): 291-4.

Alder, B. J. and T. E. Wainwright (1959). "Studies in Molecular Dynamics. I. General Method." The Journal of Chemical Physics **31**(2): 459-466.

Bahadur, R. P., P. Chakrabarti, et al. (2004). "A dissection of specific and non-specific protein-protein interfaces." J Mol Biol **336**(4): 943-55.

Bekker, H., E. Dijkstra, et al. (1995). "An efficient box shape independent non-bonded force and virial algorithm for molecular dynamics." Molecular Simulation **14**(3): 137-152.

Bennet, C. M. (1976). J. Comput. Chem. **22**: 245-268.

Berendsen, H. J., J. P. M. Postma, et al. (1984). "Molecular Dynamics with coupling to an external bath." J Chem Phys **81**: 3684-3690.

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

Borgatti, S. (1994). "How to explain Hierachical Clustering." Connections **17**(2): 78-80.

Bryngelson, J. D., J. N. Onuchic, et al. (1995). "Funnels, pathways, and the energy landscape of protein folding: a synthesis." Proteins **21**(3): 167-95.

Bryngelson, J. D. and P. G. Wolynes (1989). "Intermediates and barrier crossing in a random energy model. ." Journal of Physical Chemistry **93**: 6902-6915.

Buckle, A. M., G. Schreiber, et al. (1994). "Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution." Biochemistry **33**(30): 8878-89.

Bussi, G., D. Donadio, et al. (2007). "Canonical sampling through velocity rescaling." J Chem Phys **126**(1): 014101.

Camacho, C. J., Z. Weng, et al. (1999). "Free energy landscapes of encounter complexes in protein-protein association." Biophys J **76**(3): 1166-78.

Ceruso, M. A., X. Periole, et al. (2004). "Molecular dynamics simulations of transducin: interdomain and front to back communication in activation and nucleotide exchange." J Mol Biol **338**(3): 469-81.

Chen, R., L. Li, et al. (2003). "ZDOCK: an initial-stage protein-docking algorithm." Proteins **52**(1): 80-7.

Chen, R., J. Mintseris, et al. (2003). "A protein-protein docking benchmark." Proteins **52**(1): 88-91.

Chen, R. and Z. Weng (2002). "Docking unbound proteins using shape complementarity, desolvation, and electrostatics." Proteins **47**(3): 281-94.

Chothia, C. and J. Janin (1975). "Principles of protein-protein recognition." Nature **256**(5520): 705-8.

Comeau, S. R., D. W. Gatchell, et al. (2004). "ClusPro: a fully automated algorithm for protein-protein docking." Nucleic Acids Res **32**(Web Server issue): W96-9.

Comeau, S. R., D. W. Gatchell, et al. (2004). "ClusPro: an automated docking and discrimination method for the prediction of protein complexes." Bioinformatics **20**(1): 45-50.

Cornell, W., P. Cieplak, et al. (1995). "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." J. Am. Chem. Soc. **117**: 5179-5197.

D'andrade, R. (1978). "U-Statistic Hierarchical Clustering " Psychometrika **4**: 58-67.

Darden, T., D. York, et al. (1993). "Particle Mesh Ewald: An N-log(N) method for Ewald sums in large systems." J. Chem. Phys. **98**: 10089-10092.

Dill, K. A. (1990). "Dominant forces in protein folding." Biochemistry **29**(31): 7133-55.

Dill, K. A. (1999). "Polymer principles and protein folding." Protein Sci **8**(6): 1166-80.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.

Gabb, H. A., R. M. Jackson, et al. (1997). "Modelling protein docking using shape complementarity, electrostatics and biochemical information." J Mol Biol **272**(1): 106-20.

Gabdoulline, R. R. and R. C. Wade (1999). "On the protein-protein diffusional encounter complex." J Mol Recognit **12**(4): 226-34.

Gabdoulline, R. R. and R. C. Wade (2001). "Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations." J Mol Biol **306**(5): 1139-55.

Gabdoulline, R. R. and R. C. Wade (2002). "Biomolecular diffusional association." Curr Opin Struct Biol **12**(2): 204-13.

Ghosh, M., G. Meiss, et al. (2007). "The nuclease a-inhibitor complex is characterized by a novel metal ion bridge." J Biol Chem **282**(8): 5682-90.

Gray, J. J., S. Moughon, et al. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." J Mol Biol **331**(1): 281-99.

Grossfield, A. (2003). "An Implementation of WHAM: The Weighted Histogram Analysis Method."

Guex, N. and M. C. Peitsch (1997). "SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling." Electrophoresis **18**: 2714-2723.

Hartley, R. W. (1993). "Directed mutagenesis and barnase-barstar recognition." Biochemistry **32**(23): 5978-84.

Hartley, R. W., V. Both, et al. (1996). "Barstar Inhibits Extracellular Ribonucleases of Streptomyces and Allows their Production from Recombinant Genes." Protein Pept. Lett. **3**: 225-248.

Hastie, T., R. Tibshirani, et al. (2009). "The Elements of Statistical Learning." Springer, New York, NY, USA: 175.

Hockney, R., S. Goel, et al. (1974). "Quiet high-resolution computer models of a plasma." Journal of Computational Chemistry **14**(2).

Hoefling, M. and K. E. Gottschalk (2010). "Barnase-Barstar: From first encounter to final complex." J Struct Biol **171**(1): 52-63.

Hornak, V., R. Abel, et al. (2006). "Comparison of multiple Amber force fields and development of improved protein backbone parameters." Proteins **65**(3): 712-25.

Hubbard, S. J. and P. Argos (1994). "Cavities and packing at protein interfaces." Protein Sci **3**(12): 2194-206.

Hubbard, S. J. and J. M. Thornton (1993). "NACCESS: A computer program." Department of Biochemistry and Molecular Biology, University College London.

Hwang, H., B. Pierce, et al. (2008). "Protein-protein docking benchmark version 3.0." Proteins **73**(3): 705-9.

Hwang, H., B. Pierce, et al. (2009). Protein - Protein Docking. Computational Protein-Protein Interactions. R. Nussinov and G. Schreiber, CRC Press**:** 147-165.

Hwang, H., T. Vreven, et al. (2010). "Protein-protein docking benchmark version 4.0." Proteins **78**(15): 3111-4.

Janin, J. (1997). "Specific versus non-specific contacts in protein crystals." Nature **4**(12): 973-974.

Janin, J. (1997). "The kinetics of protein-protein recognition." Proteins **28**(2): 153-61.

Janin, J. (2009). Basic Principles of Protein-Protein Interactions. Computational Protein-Protein Interactions. R. Nussinov and G. Schreiber, CRC Press**:** 1-19.

Janin, J. and C. Chothia (1990). "The structure of protein-protein recognition sites." J Biol Chem **265**(27): 16027-30.

Jiang, S., A. Tovchigrechko, et al. (2003). "The role of geometric complementarity in secondary structure packing: a systematic docking study." Protein Sci **12**(8): 1646-51.

Johnson, S. (1967). "Hierachical Clustering Schemes " Psychometrika **2**: 241-254.

Jones, S. and J. M. Thornton (1996). "Principles of protein-protein interactions." Proc Natl Acad Sci U S A **93**(1): 13-20.

Jones, S. and J. M. Thornton (1997). "Prediction of protein-protein interaction sites using patch analysis." J Mol Biol **272**(1): 133-43.

Jorgensen, W., D. Maxwell, et al. (1996). "Developement and testing of the OPLS all-atom force field on conformational energies and properties of organic liquids." Journal of the American Chemical Society **118**: 11225-11236.

Jorgensen, W. L. and J. Tirado-Rives (1988). "The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin." J. Am. Chem. Soc. **110**: 1657-1666.

Karplus, M. and A. Sali (1995). "Theoretical studies of protein folding and unfolding." Curr Opin Struct Biol **5**(1): 58-73.

Kastritis, P. L., I. H. Moal, et al. (2011). "A structure-based benchmark for protein-protein binding affinity." Protein Sci **20**(3): 482-91.

Katchalski-Katzir, E., I. Shariv, et al. (1992). "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques." Proc Natl Acad Sci U S A **89**(6): 2195-9.

Keskin, O., C. J. Tsai, et al. (2004). "A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications." Protein Sci **13**(4): 1043-55.

Kirkwood, J. (1935). "Statistical mechanics of fluid mixtures." J. Chem. Phys. **3**: 300.

Korn, A. P. and R. M. Burnett (1991). "Distribution and complementarity of hydropathy in multisubunit proteins." Proteins **9**(1): 37-55.

Kozlov, G., L. Nguyen, et al. (2007). "Structural basis of ubiquitin recognition by the ubiquitin-associated (UBA) domain of the ubiquitin ligase EDD." J Biol Chem **282**(49): 35787-95.

Kozlov, G., P. Peschard, et al. (2007). "Structural basis for UBA-mediated dimerization of c-Cbl ubiquitin ligase." J Biol Chem **282**(37): 27547-55.

Kumar, S., D. Bouzida, et al. (1992). "The Weighted Histogram Analysis Method for Free Energy Calculations on Biomolecules. I. The Method." J. Comput. Chem. **13**: 1011-1021.

Lawrence, M. C. and P. M. Colman (1993). "Shape complementarity at protein/protein interfaces." J Mol Biol **234**(4): 946-50.

Lazaridis, T. and M. Karplus (1997). ""New view" of protein folding reconciled with the old through multiple unfolding simulations." Science **278**(5345): 1928-31.

Lyskov, S. and J. J. Gray (2008). "The RosettaDock server for local protein-protein docking." Nucleic Acids Res **36**(Web Server issue): W233-8.

MacKerell, J. A. D., et al (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." J Chem Phys **102**: 3586-3616.

Marrink, S. J., A. H. deVries, et al. (2004). "Coarse Grained Model for Semiquantitative Lipid Simulations." J. Phys. Chem. B **108**(2): 750-760.

Marrink, S. J., H. J. Risselada, et al. (2007). "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations." J. Phys. Chem. B **111**(27): 7812-7824.

Monticelli, L., S. K. Kandasamy, et al. (2008). "The MARTINI Coarse-Grained Force Field: Extension to Proteins." J. Chem. Theory Comput. **4**(5): 819-834.

Neumann, J. and K. E. Gottschalk (2009). "The effect of different force applications on the protein-protein complex Barnase-Barstar." Biophys J **97**(6): 1687-99.

Ofran, Y. (2009). Prediction of Protein Interaction Sites. Computational Protein-Protein Interactions. R. Nussinov and G. Schreiber, CRC Press**:** 167-184.

Ofran, Y. and B. Rost (2007a). "ISIS: interaction sites identified from sequence." Bioinformatics **23**(2): e13-6.

Ofran, Y. and B. Rost (2007b). "Protein-protein interaction hotspots carved into sequences." PLoS Comput Biol **3**(7): e119.

Pachov, G. V., R. R. Gabdoulline, et al. (2009). Computational Simulations of Protein - Protein and Protein - Nucleic Acid Association. Computational Protein-Protein Interactions. R. Nussinov and G. Schreiber, CRC Press**:** 109-127.

Papoian, G. A. and P. G. Wolynes (2003). "The physics and bioinformatics of binding and folding-an energy landscape perspective." Biopolymers **68**(3): 333-49.

Periole, X., M. Cavalli, et al. (2009). "Combining an Elastic Network with a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition " J. Chem. Theory Comput. **5**: 2531 - 2543.

Peschard, P., G. Kozlov, et al. (2007). "Structural basis for ubiquitin-mediated dimerization and activation of the ubiquitin protein ligase Cbl-b." Mol Cell **27**(3): 474-85.

Ponder, J. W. and F. M. Richards (1987). "Internal packing and protein structural classes." Cold Spring Harb Symp Quant Biol **52**: 421-8.

Richards, F. M. (1977). "Areas, volumes, packing and protein structure." Annu Rev Biophys Bioeng **6**: 151-76.

Rizzi, R., P. Mahata, et al. "Hierarchical clustering using the arithmetic-harmonic cut: complexity and experiments." PLoS One **5**(12): e14067.

Roux, B. (1995). "The calculation of the potential of mean force using computer simulations." Computer Physics Communications **91**: 275 - 282.

Rual, J. F., K. Venkatesan, et al. (2005). "Towards a proteome-scale map of the human protein-protein interaction network." Nature **437**(7062): 1173-8.

Schreiber, G. (2002). "Kinetic studies of protein-protein interactions." Curr Opin Struct Biol **12**(1): 41-7.

Schreiber, G. (2009). The Association of Protein - Protein Complexes. Computational Protein-Protein Interactions. R. Nussinov and G. Schreiber, CRC Press**:** 87-107.

Schreiber, G. and A. R. Fersht (1993). "Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering." Biochemistry **32**(19): 5145-50.

Schreiber, G. and A. R. Fersht (1996). "Rapid, electrostatically assisted association of proteins." Nat Struct Biol **3**(5): 427-31.

Selzer, T. and G. Schreiber (2001). "New insights into the mechanism of protein-protein association." Proteins **45**(3): 190-8.

Sevcik, J., L. Urbanikova, et al. (1998). "Recognition of RNase Sa by the inhibitor barstar: structure of the complex at 1.7 A resolution." Acta Crystallogr D Biol Crystallogr **54**(Pt 5): 954-63.

Sorin, E. J. and V. S. Pande (2005). "Exploring the helix-coil transition via all-atom equilibrium ensemble simulations." Biophys J **88**(4): 2472-93.

Spaar, A., C. Dammer, et al. (2006). "Diffusional encounter of barnase and barstar." Biophys J **90**(6): 1913-24.

Spaar, A. and V. Helms (2005). "Free energy landscape of protein-protein encounter resulting from Brownian Dynamics Simulations of Barnase:Barstar." Journal of Chemical Theory and Computation **1**: 723-736.

Strittmatter, G., J. Janssens, et al. (1995). "Inhibition of Fungal Disease Developement in Plants by Engineering Controlled Cell Death." Biotechnology **13**: 1985-1089.

Tirion, M. M. (1996). "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." Phys Rev Lett **77**(9): 1905-1908.

Torrie, G. M. and J. P. Valleau (1977). "Nonphysical Sampling Distributions in Monte Carlo Free Energy Estimation: Umbrella Sampling." J. Chem. Phys. **23**: 187-199.

Tozzini, V. (2005). "Coarse-grained models for proteins." Curr Opin Struct Biol **15**(2): 144-50.

Tsai, C. J., S. Kumar, et al. (1999). "Folding funnels, binding funnels, and protein function." Protein Sci **8**(6): 1181-90.

Tsai, C. J., S. L. Lin, et al. (1996). "Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences." Crit Rev Biochem Mol Biol **31**(2): 127-52.

Tsai, C. J., B. Ma, et al. (1999). "Folding and binding cascades: shifts in energy landscapes." Proc Natl Acad Sci U S A **96**(18): 9970-2.

Van Der Spoel, D., E. Lindahl, et al. (2005). "GROMACS: fast, flexible, and free." J Comput Chem **26**(16): 1701-18.

Van Der Spoel, D. and P. J. Van Maaren (2006). "The Origin of Layer Structure Artifacts in Simulations of Liquid Water." Journal of Chemical Theory and Computation **2**(1): 1-11.

Van Gunsteren, W. F., S. R. Billeter, et al. (1996). "Biomolecular simulation: The GROMOS96 manual and user guide. ." Zürich, Switzerland: Hochschulverlag AG an der ETH Zürich.

Verlet, L. (1967). "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard−Jones Molecules." Phys. Rev. **159**(98-103): 98.

Wang, J., P. Cieplak, et al. (2000). "How well does a Restrained Electrostatic Potential (RESP) model perform in calculating conformational energies of organic and biological molecules." J. Comput. Chem. **21**(12): 1049-1074.

Wang, L., S. W. Siu, et al. (2010). "Downhill binding energy surface of the barnase-barstar complex." Biopolymers **93**(11): 977-85.

Zhang, C., J. Chen, et al. (1999). "Protein-protein recognition: exploring the energy funnels near the binding sites." Proteins **34**(2): 255-67.