Molecular Modeling of Biomolecular Associations and Quantifying Allosteric Effects

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich–Technischen Fakultät III

Chemie, Pharmazie, Bio- und Werkstoffwissenschaften

der Universität des Saarlandes

von

Özlem Ulucan

Saarbrücken

April 2015

Tag des Kolloquiums:	15.06.2015
Dekan:	Prof. Dr.-Ing. Dirk Bähre
Berichterstatter:	Prof. Dr. Volkhard Helms
	Prof. Dr. Richard Zimmermann
Vorsitz:	Prof. Dr. Karin Römisch
Akad. Mitarbeiter:	Dr. Matthias Engel

Abstract

Molecular dynamics simulation technique is a very popular approach to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes. Using extensive standard molecular dynamics simulations and variants thereof, we probed in this work structural and energetic aspects of either specific or non-specific protein-protein complexes formed by hydrophilic proteins and of the interfacial water between the two proteins. For the specific complexes, the standard free energies of binding are in overall good agreement with the experimental values. In comparison to their specific counterparts, non-specific encounters bear smaller interaction interfaces and are attracted by shorter-ranged direct interactions between the proteins.

In order to quantify the allosteric effect, we calculated the allosteric coupling energy between the ATP binding pocket and the PIF-pocket of phosphoinostide-dependent kinase-1. For this system, we found that the main contribution to the allosteric coupling energy comes from electrostatic interactions.

Utilizing molecular docking we modeled the interaction of the ER luminal binding immunoglobulin protein (BIP) with loop 7 of the Sec61 translocon. Additionally we used a cyclic peptide as a scaffold to design new competitive compounds that bind to Casein kinase II α in a competitive manner to Casein kinase II β using molecular docking.

Kurzfassung

Die molekulardynamische Simulationstechnik ist heute weitverbreit, um die Struktur, Dynamik und Thermodynamik von Biomolekülen und ihren Komplexen zu untersuchen. Mithilfe umfangreicher gewöhnlicher sowie spezieller MD-Simulationstechniken untersuchten wir die strukturellen und energetischen Aspekte von spezifischen oder nicht-spezifischen Protein-Protein-Komplexen, die durch hydrophile Proteine gebildet werden, sowie für das Wasser im Zwischenraum zwischen den beiden Molekülen. Für die spezifischen Komplexe sind die Standard-Gibbs-Bindungsenergien insgesamt in guter Übereinstimmung mit den experimentellen Werten. Im Vergleich zu ihren spezifischen Pendants werden unspezifische Proteinkontakte durch kleinere Interaktionsschnittstellen vermittelt, und die attraktiven Wechselwirkungen zwischen den Proteinen haben eine kürzere Reichweite.

Um allosterische Effekte zu quantifizieren, berechneten wir die allosterische Kopplungsenergie zwischen der ATP-Bindungstasche und der PIF-Tasche des Phosphoinostid-abhängigen Kinase-1. Für dieses System stellten wir fest, dass der Hauptbeitrag zu der allosterischen Kopplungsenergie von elektrostatischen Wechselwirkungen kommt.

Durch molekulares Docking modellierten wir die Wechselwirkung des ER luminalen Proteins BIP mit der Schlaufe 7 des Sec61-Translokons. Mithilfe von molekularem Docking konstruierten wir zudem ein zyklisches Peptidgerüst für neue Casein Kinase II β (CK2β) kompetitive Peptide, die an CK2α binden.

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Volkhard Helms for giving me the opportunity to work in his group and for his guidance and support throughout this thesis. Your profound knowledge and experience in biomolecular simulations have been of great help for my professional development. Many thanks for your constructive suggestions, fruitful discussions, grateful encouragement and all other contributions that I cannot list here.

Work would not have been as pleasant without my colleagues. I would like to thank the entire Helms group. Special thanks to Peter, Wei, and Mazen for helping me to develop my skills and providing their kind friendship any time. I would like to thank our secretary Kerstin for helping me to overcome many bureaucratic hurdles with visa extensions and other administration issues.

I would also like to thank my roommates Cinthia, Rachel and Linda and my office mates Siti, Mazen and Siba and my friend and colleague Mohamed. Thanks for the shared moments. Having you in my life was a great chance for me. I also owe special thanks to Ahmad for helping me during the submission process and for the delicious maqluba we had all together.

I would like to show my gratitude to Prof. Dr. Richard Zimmermann, Dr. Matthias Engel and Dr. Karsten Niefind for the joint projects that we carried out together. Collaborations with you were instructive and exiting.

I would like to gratefully acknowledge the Deutsche Forschungsgemeinschaft for the financial support and the Gauss Center for Supercomputing for providing computing time on the supercomputer SuperMUC at Leibniz Supercomputer Centre.

I owe my deepest gratitude to my parents who was always there for me. Thank you for your unconditional support!

My special thanks goes to my husband Fesih for his unflagging love and support.

List of Figures

List of Tables

Contents

1. Introduction

In biological systems, molecules are continually binding together and dissociating from each other. Molecular associations are the first steps in many biological processes such as the cell cycle, cellular transport, immune response, apoptosis, DNA replication and transcription, RNA splicing and signal transduction. Several conceptual models have been proposed in the literature to explain the mechanism of molecular associations. In 1894, Emil Fischer attributed the extraordinary specificity of enzymes to a strict geometric complementarity between enzymes and their substrates. Later, this idea became famous as "lock-and-key" model to explain the recognition between molecules and proteins. In 1958, Daniel Koshland from the University of California, Berkley, recognized that proteins are flexible structures and hypothesized that they accommodate the substrate by adjusting their shapes around it. Likewise, the substrate could adapt its shape, as well, to optimize interactions with its binding partner. This concept is nowadays referred to as "induced fit" hypothesis. A further model was introduced that is termed "conformational selection", which describes the stabilization of global protein conformations through ligand binding. According to this model, a protein samples in its unbound form a multitude of interconvertable conformational states. The binding of a ligand initiates a process of conformational selection within the ensemble where the ligand preferentially binds to the conformations of the protein for which it has the highest affinity at the expense of others. Therefore, the ligand does not create new protein states but causes a shift in the population in the favor of the binding-competent state(s). There are two guiding principles in understanding association processes in molecular biology[1]. (1) The forces that govern association are almost always noncovalent. In other words, noncovalent interactions such as electrostatic forces, hydrogen bonds, and hydrophobic interactions combine in various ways to stabilize molecular complexes. (2) Associations are stereospecific and depend on a precise spatial arrangement of the interacting groups. As a result, biological associations are highly specific so

that molecules can recognize one another and distinguish subtle variations in structure.

The importance of molecular interactions such as protein-protein interactions has been realized long ago. However, it took until the 1990ies to introduce several important concepts on how they are formed. At the end of the 1980s, the "new view" of protein folding was established that describes the folding of peptide chains as a diffusive process in an energy funnel[2]. In analogy to protein folding, the related process of protein-protein binding has then also been portrayed by such a funnel[3]. Both the protein folding and protein-protein binding processes are driven by the decrease in total Gibbs free energy (at constant temperature and pressure) of the protein-solvent or protein-protein-solvent system, which is dictated by opposing effects between entropy and enthalpy, eventually leading to a global free energy minimum of these thermodynamic systems. The requirement for lowering the free energy while reducing conformational space ensures that the energy landscape of a foldable protein is funnel-shaped[2, 4]. The spontaneous protein-protein association also lowers the free energy of the system composed of proteins and solvent, while reducing the conformational, rotational and translational entropy of the binding partners[3]. Accordingly, the binding free energy is the key driving force behind protein-protein association and of fundamental importance in supramolecular chemistry.

Water molecules play an invaluable role in governing the structure, stability, dynamics, and function of biomolecules. However, the exact range of processes mediated by water is far from being understood. In many natural systems, water is confined in an environment where its free movement is restricted and its three-dimensional hydrogen-bonded network is partially disrupted. The properties of such confined water are difficult to predict and may be considerably different from those of bulk water. Hydrophobic dewetting has been reported as a general mechanism for the association of hydrophobic surfaces[5]. However, for the assembly of hydrophilic surfaces, the picture looks a bit more complicated.

Decades of experimental and theoretical efforts have led to the development of paradigms, theoretical methods and computational tools to study molecular

interactions. Molecular modeling techniques are heavily used, for instance, in computational chemistry, computational biology, drug discovery and material science. Some of the basic molecular modeling techniques are quantum mechanical and molecular mechanics methods, Brownian and molecular dynamics simulations and molecular docking. Molecular dynamics is the principal tool for theoretical modeling of the conformational dynamics and energetics of biological molecules. This computational method is now routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes.

The main goal of this thesis is to study structural and energetics aspects of specific and non-specific protein-protein complexes formed by hydrophilic proteins in a comparative manner. To do this, we utilized several theoretical approaches for analyzing trajectories obtained from standard and variant molecular dynamics simulations.

This thesis consists of seven chapters. In the present chapter, we convey an overview of molecular interactions and their importance for biological systems. In the second chapter, theoretical aspects of the used concepts and methods are thoroughly introduced. Chapters 3, 4, 5 and 6 are organized in a similar manner. Each of those chapters describes a complete project with the background, the methods used and the results obtained combined with the discussion section.

Chapter 3 addresses the association of hydrophilic proteins to form specific complexes and the role of water that were studied by means of extensive molecular dynamics simulations in explicit water. The systems of interest are the three well studied protein complexes; Barnase - Barstar, Cytochrome *c* – Cytochrome *c* peroxidase, and the N-terminal domain of enzyme I – Histidine-containing Phosphocarrier. One-dimensional free energy profiles of association were obtained using umbrella-sampling simulations. Properties of the interfacial water localized between two complex partners were quantified, as well.

In Chapter 4, using the analogues methods utilized in Chapter 3, we probed comprehensively the structural and energetics aspects of nonspecific complexes formed by the same protein pairs. In that chapter, we were seeking for the source of interaction specificity that favors a small set of interactions over the multitude of possibilities.

In Chapter 5, we attempted to model the universal phenomenon of allostery with a thermodynamic two-state model. We present the results from alchemical free energy perturbation calculations to quantify the allosteric affect in a case study of phosphoinostide-dependent kinase-1 (PDK1).

Chapter 6 discusses applications of the widely used docking program AutoDock to study aspects of protein-protein interaction. After introducing the docking algorithms and scoring functions employed, we present two fruitful applications that we performed in collaboration with experimental groups.

In the last chapter we summarize our work and its main findings followed by an outlook others can build on it.

2. Binding Free Energy Theory and Computation

2.1. Statistical Thermodynamics of Binding

2.1.1. Thermodynamics and Standard State

Conceptually, one can think of a binding reaction as a noncovalent association of two molecules A and B that form the complex AB[6];

$$A + B \rightleftarrows AB \quad (2.1)$$

In the case of biological systems, the above reaction typically takes place in a mixed solvent, which is called reaction solvent. The condition for equilibrium is;

$$\mu_{sol,A} + \mu_{sol,B} = \mu_{sol,AB} \quad (2.2)$$

where $\mu_{sol,i}$ is the chemical potential of species $i = A, B$ or AB in solution.

The chemical potential of species i in solution is given by the following equation:

$$\mu_{sol,i} = \mu^{o}_{sol,i} + RTln\frac{\gamma_i C_i}{C^o} \quad (2.3)$$

In this equation $\mu^{o}_{sol,i}$ and C_i are the standard chemical potential and the concentration of species i, respectively. R is the gas constant, T is the absolute temperature, γ_i is the activity coefficient of i and C^o is the standard concentration in the same units as C_i. It is common to express the concentration in molar units (M), in which case $C^o = 1M$. Frequently, C^o is not included explicitly. In such cases, C_i is then the ratio of the concentration over the standard concentration.

The standard chemical potential of species A, B and AB is their chemical potential in standard state conditions. It is strategical to define a hypothetical state in which each species is at standard state in the reaction solvent but does not interact with other molecules of A, B or AB[6, 7]. The activity coefficients relative to this standard state, γ_i, approach unity as C_i approaches zero in the reaction solvent.

Considering the above equations, the standard free energy of binding can be calculated as follows:

$$\Delta G^o_{AB} = \mu^o_{sol,AB} - \mu^o_{sol,A} - \mu^o_{sol,B} \quad (2.4)$$

$$\Delta G^o_{AB} = -RTln\left(\frac{\gamma_{AB}}{\gamma_A\gamma_B}\frac{C^oC_{AB}}{C_AC_B}\right)_{eq} \quad (2.5)$$

$$\Delta G^o_{AB} = -RTlnK_{AB} \quad (2.6)$$

where $(\quad)_{eq}$ implies a quantity at equilibrium. In the case of low concentrations, $\gamma_i = 1$ is typically a good approximation. Since C^o is not written explicitly in expressions for the binding constant K_{AB}, it may appear that K_{AB} bears units of volume. However, K_{AB} is a dimensionless quantity. When C^o is not written explicitly, it is implicit that the units of concentration are standard concentration.

2.1.2. The Standard Chemical Potential of a Molecule in Solution

The standard chemical potential of a molecule A $(\partial G/\partial n_A)_{T,P}$ in solution is given by:

$$\mu^o_{sol,A} = -RTln\left(\frac{1}{V_{N,A}C^o}\frac{Q_{N,A}(V_{N,A})}{Q_{N,O}(V_{N,O})}\right) + P^o\bar{V}_A \quad (2.7)$$

For the detailed derivation of the above equation, please see[8]. $Q_{N,A}(V_{N,A})$ is the canonical partition function for a system containing N solvent molecules and one solute molecule A at volume $V_{N,A}$, which is the volume of the system when it is at equilibrium at standard state pressure P^o(1 atmosphere). Likewise, $Q_{N,O}(V_{N,O})$ is the partition function for the N solvent molecules without solute, at a different equilibrium volume $V_{N,O}$ that also corresponds to the pressure P^o. $\bar{V}_A = V_{N,A} - V_{N,O}$ is the change in equilibrium volume when one molecule of solute is added to the N solvent molecules. Thus, for the case of $N \gg 1$, $\bar{V}_A$ is the partial molar volume of the solute at infinite dilution in the solvent [8]. The above expression for $\mu^o_{sol,A}$ can be interpreted as the standard chemical potential of the solute in the gas phase added to the work of transferring it to the solvent under constant pressure.

The energy of the system in terms of conjugate momenta and coordinates is given as follows:

$$H(p_A, p_S, r'_A, r_S) = \sum_{i=1}^{M_A+M_S} \frac{p_i^2}{2m_i} + U(r'_A, r_s) \quad (2.8)$$

Here, M_A and M_B are respectively the number of atoms of the single solute molecule and of the N solvent molecules; p is a vector of the momenta of the $M_A + M_S$ atoms, p_i^2 is the squared magnitude of the momentum of atom i; m_i is the mass of atom i; and $U(r_A', r_s)$ is the potential energy as a function of all the atomic coordinates. Assuming that classical statistical mechanics is applicable we get:

$$\frac{Q_{N,A}}{Q_{N,O}} = \frac{\int dp_A dp_s \int dr_A' dr_S \, e^{-\beta\left(\sum_{i=1}^{M_A+M_S} p_i^2/2m_i + U(r_A', r_S)\right)}}{\sigma_A \int dp_s \int dr_S \, e^{-\beta\left(\sum_{i=M_A+1}^{M_A+M_S} p_i^2/2m_i + U(r_S)\right)}} \quad (2.9)$$

where $\beta = (RT)^{-1}$ and σ_A is the symmetry number of the solute. Atoms $i = 1, \ldots M_A$ belong to the solute, atoms $i = M_A + 1, \ldots M_A + M_S$ belong to the solvent.

The lab-frame coordinates of the solute atoms r_A' can be separated into internal and external coordinates based on a molecular axis system. Any three atoms (atom 1, 2 and 3) may be used to define the molecular axes. The Cartesian coordinate of each atom may be specified relative to the newly defined molecular axes. Note that, in internal coordinates, atom 1 is fixed at the origin; atom 2 always lies on the x-axis; and atom 3 lies in the $z = 0$ plane. The six coordinates thus fixed correspond to the external coordinates of the molecule. We will term the $3M_A - 6$ coordinates as r_A and the complete set of external coordinates as ζ_A. The integrals over the internal coordinates of the solvent and as well as of the solute do not depend upon the position or orientation of the solute. Therefore the integral over ζ_A yields a factor of $8\pi^2 V_{N,A}$. The integral over the momentum for each atom i yields a factor of $(2\pi m_i RT)^{3/2}$ in the classical approximation. The momentum integrals for the solvent atoms cancel. Therefore, the standard chemical potential of species A is[6]:

$$\mu_{sol,A}^o = -RTln\left(\frac{8\pi^2}{C^o \sigma_A}\prod_{i=1}^{M_A}(2\pi m_i RT)^{3/2}\frac{Z_{N,A}}{Z_{N,O}}\right) + P^o \bar{V}_A \quad (2.10)$$

where $Z_{N,A}$ and $Z_{N,O}$ are configuration integrals and are given by:

$$Z_{N,A} = \int e^{-\beta U(r_A, r_S)} dr_A dr_S \quad (2.11)$$

$$Z_{N,O} = \int e^{-\beta U(r_S)} dr_S \quad (2.12)$$

2.1.3. The Standard Chemical Potential of a Molecular Complex in Solution

As for the single molecule, one has to define internal and external coordinates of the complex. Using the external coordinates of molecule A as the external coordinates of the complex is one way of doing that. Then the external coordinates of B, ζ_B, are defined relative to molecule A. Therefore, the external coordinates of B become internal coordinates of the complex.

It is clear that one should not include the configurations for which the two molecules are far apart in the configurational integral of the complex. This can be accomplished by introducing a step function $I(\zeta_B)$ that is equal to unity for the bound configurations and zero otherwise.

Thus the standard chemical potential of the complex is[6]:

$$\mu^o_{sol,AB} = -RTln\left(\frac{8\pi^2}{C^o\sigma_{AB}} \prod_{i=1}^{M_A+M_B} (2\pi m_i RT)^{3/2} \frac{Z_{N,AB}}{Z_{N,O}}\right) + P^o\bar{V}_{AB} \quad (2.13)$$

$$Z_{N,AB} = \int I(\zeta_B) J_{\zeta_B} e^{-\beta U(r_A, r_B, \zeta_B, r_S)} dr_A dr_B d\zeta_B dr_S \quad (2.14)$$

Here $M_A + M_B$ is the number of atoms in both solutes, J_{ζ_B} is the Jacobian determinant for the Eulerian rotation of molecule B with respect to molecule A.

It is generally not straightforward to define the complex, meaning the range over which $I(\zeta_B)$ equals unity. Particularly simple is defining $I(\zeta_B)$ whenever the potential of mean force for the interaction of A and B is sharply peaked and negative in a small range of ζ_B. In such a case the stable configurations of the complex will dominate the thermodynamic averages over the zone of configuration space where A and B are close together. According to Hill[9], the region in which $I(\zeta_B)$ is equal to one must satisfy the following two criteria:

1. The region should include all the configurations that contribute significantly to the chemical potential of the bound state; namely those for which the Boltzmann factor of the potential of mean force is large.
2. The region should not include so large a phase volume of unbound configurations those contribute appreciably to $\mu^o_{sol,AB}$. For instance, the

complex cannot be considered as formed in a configuration where A and B are infinitely far apart.

2.1.4. The Standard Free Energy of Binding

Above, we derived the standard chemical potential for a molecule and for a complex in solution. Based on these equations, we rewrite the expression for the standard free energy of binding in terms of molecular properties[6]:

$$\Delta G^{o}_{AB} = -RTln\left(\frac{C^{o}}{8\pi^{2}}\frac{\sigma_A\sigma_B}{\sigma_{AB}}\frac{Z_{N,AB}}{Z_{N,A}}\frac{Z_{N,O}}{Z_{N,B}}\right) + P^{o}\Delta\bar{V}_{AB} \quad (2.15)$$

where $\Delta\bar{V}_{AB} = \bar{V}_{AB} - \bar{V}_A - \bar{V}_B$. As seen in the above equation, all mass dependent terms have cancelled.

2.2. Molecular Dynamics

Molecular dynamics (MD) is a type of computer simulation in which particles are allowed to interact for a period of time using approximations of classical mechanics. MD is an interface between laboratory experiments and theory, and is, therefore, a *"virtual experiment"*. MD inquires the relationship between molecular structure, movement and function.

The molecular dynamics simulation method is based on Newton's second law known as equation of motion, $F = ma$, where F is the force exerted on the particle, m and a are the mass and the acceleration of the particle, respectively. If the forces acting on each atom are known, it is straightforward to determine the acceleration of each atom in the system of interest. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time. MD is a deterministic method meaning that, once the positions and velocities of each atom are known, the state of the system can be predicted at any time.

Newton's equation of motion is given by;

$$F_i = m_i a_i \quad (2.16)$$

where F_i is the force acting on particle i, m_i is the mass of particle i and a_i is the acceleration of particle i. The force can also be expressed as the gradient of the potential energy as follows

$$F_i = -\nabla_i V \quad (2.17)$$

Combining these two equations yields

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \quad (2.18)$$

where V is the potential energy of the system and r_i is the position of atom i. Thus, the Newton's equations of motion relate the derivatives of the potential energies to the changes in positions as a function of time. In summary, to calculate a trajectory, one only needs the initial positions of the atoms and an initial distribution of velocities. The accelerations are then determined by the gradients of the potential energy functions. The equations of motion are deterministic, e.g., the positions and the velocities at time $t = 0$ determine the positions and velocities at all other times, t. The initial positions can be obtained from experimental structures, such as the X-ray crystal structures of macromolecules or their solution structures determined by NMR spectroscopy. The initial distribution of velocities is usually determined from a random distribution with the magnitudes conforming to the required temperature and ensuring that there exists no overall momentum

$$P = \sum_{i=1}^{N} m_i v_i = 0 \quad (2.19)$$

The velocities v_i are often chosen randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature. This gives the probability that an atom i has a velocity v_{ix}, in the x direction at a temperature T.

$$P(v_{ix}) = \left(\frac{m_i}{2\pi k_\beta T}\right)^{1/2} e^{-m_i v_{ix}^2 / 4 k_\beta T} \quad (2.20)$$

Methods for solving Newton's equations of motion are generally called Verlet methods named after L. Verlet, one of the pioneers who applied integration algorithms to molecular simulations. The Verlet algorithm[10] can be derived by writing two third-order Taylor expansions for the positions $r(t)$, one forward and one backward in time. Denoting the velocities with v, the accelerations with a and the third derivatives of r with respect to t with b, one obtains:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \frac{1}{6}b(t)\Delta t^3 + O(\Delta t^4) \quad (2.21)$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 - \frac{1}{6}b(t)\Delta t^3 + O(\Delta t^4) \quad (2.22)$$

Summing up the expressions gives:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \quad (2.23)$$

which is the basic form of the algorithm. An issue with this basic form of the Verlet algorithm is that it does not explicitly generate velocities. Even though they are not needed for the time evolution, their knowledge is sometimes necessary, for instance for the calculation of the particles' kinetic energies and thus the system temperature. To overcome this difficulty, some variants of the Verlet algorithm have been developed, such as the leap-frog algorithm[11] and the velocity Verlet scheme. The leap-frog algorithm uses forces $F(t)$ and positions $r(t)$ at time t and the velocities $v(t-\frac{\Delta t}{2})$ at time $(t-\frac{\Delta t}{2})$ to update the positions and velocities.

$$v\left(t+\frac{\Delta t}{2}\right) = v\left(t-\frac{\Delta t}{2}\right) + \frac{\Delta t}{m}F(t) \quad (2.24)$$

$$r(t+\Delta t) = r(t) + \Delta t v\left(t+\frac{\Delta t}{2}\right) \quad (2.25)$$

2.2.1. Modeling Interactions in Molecular Dynamics

In molecular dynamics simulations, the interactions between particles are modeled by molecular mechanics force fields. These force fields are based on the Born-Oppenheimer approximation[12] to the Schrödinger equation. The Born-Oppenheimer approximation states that since the masses of the nuclei are much larger than the masses of the electrons, the electronic wave function depends only on the positions of the nuclei and not on their momenta. Therefore, the movements of electrons and atoms can be treated separately. This enables representation of atoms as classical point particles that follow classical Newtonian dynamics. In classical molecular mechanics the effect of the electrons is approximated by an effective potential function whose parameters are usually derived through fitting.

The most widely used force fields, such as AMBER[13], CHARMM[14], GROMOS[15] and OPLSAA[16], consist of two parts, the functional form and the parameters. The functional form is the description of the potential function. The potential function (U) describing interactions among particles typically comprises electrostatic, van der Waals, bond, angle, and dihedral terms and can be expressed as:

$$U = \sum_{bonds} K_b(b - b_{eq})^2 + \sum_{bond\ angles} \frac{1}{2} K_\theta(\theta - \theta_{eq}) + \sum_{dihedrals} \frac{1}{2} K_\phi[1 - \cos(n\phi + \gamma)] + \sum_{i<j} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \quad (2.26)$$

where, K_b and K_θ are the force constants for the bonds and bond angles, respectively; b and θ are the instantaneous bond lengths and bond angles; b_{eq} and θ_{eq} are the equilibrium bond lengths and bond angles; ϕ are the dihedral angles and K_ϕ is the corresponding force constant; the phase angle takes values of either 0° or 180°. The non-bonded parts of the potential are represented by van der Waals (A_{ij}), London dispersion terms (B_{ij}) and interactions between partial atomic charges (q_i and q_j). ε is the dielectric constant that accounts for effects of the medium that are not explicitly represented.

The force field parameters are most often obtained by fitting to data from experiments or from high-level quantum mechanical calculations.

The utility of MD simulations faces two main challenges: the force fields used may require further refinement, and the high computational demands of such simulations prohibit routine simulations exceeding the time scale of microseconds, leading in many cases to an inadequate sampling of conformational states. For example, a few microseconds-long MD simulation of a relatively small system may take weeks to months to complete at the current moment in time. Therefore molecular dynamics simulations are typically performed on computer clusters or supercomputers using dozens or hundreds of processors in parallel.

2.3 Approaches for Calculating Binding Affinities

2.3.1. Docking and Scoring

Docking methods try to identify the most stable bound conformation of two molecules, generally of a protein and a ligand using a simplified energy model. Often an empirical force field with a simplified solvent model is used in order to minimize the required computational cost. The conformation is then used to assign a binding energy or a score to the ligand. We will discuss docking methods in detail in chapter 6 together with some applications.

2.3.2. Free Energy Methods

Unlike docking, free energy methods utilize conformational sampling in order to generate proper thermodynamic averages. The use of conformational sampling eliminates the sensitivity to a single conformation relied on by docking methods. As a disadvantage free energy methods are computationally demanding in terms of generating statistically converged results. We will start by briefly discussing the end-point methods and then give a detailed description of pathway methods that were employed in this thesis.

2.3.2.1. End-point Methods

End-point methods generate conformations of only the free and bound species and compute the binding free energy by taking a difference. The linear interaction energy (LIE)[17, 18] and MM-PBSA[19] methods are the most popular end-point approaches.

LIE involves running two MD simulations: one for the ligand in the solution and the other for the ligand placed in the protein binding site. It is assumed that the saved snapshots represent Boltzmann ensembles of conformations and are utilized to compute the Boltzmann-averaged electrostatic and van der Waals interaction energies of the ligand with its environment in the bound and free states. Changes in the internal energy of the solvent and the protein are accounted for by some factors.

Even though LIE does not account explicitly for the standard concentration or for changes in the configurational entropy or the internal energy of the ligand, the

results may be surprisingly good if suitable scaling constants are used for the energy terms [20].

A second end-point approach MM-PBSA, along with its GB variant, relies on MD simulations of free ligand, free protein and their complex. In this approach the simulations are used for calculating the average potential and solvation energies. The snapshots that are saved during MD simulations are post processed by stripping away the explicitly modelled solvent molecules and computing their potential energies with the force field and their solvation energies with either the PBSA or GBSA implicit solvent model. The changes in the mean potential energy and solvation energy are computed by averaging over each trajectory. Then the change in configurational entropy is estimated by a rigid-rotor/harmonic-oscillator approximation using a few energy-minimized snapshots. Although it is not clear from the literature how the standard concentration has been incorporated into MM-PBSA/MM-GBSA method, ref [21] provides a clear connection of the method to the theory of binding. The main drawback of this approach is to obtain converged energy averages. This is hampered by the sizeable energy fluctuations not only of the ligand and the binding site, but also of parts of the protein remote from the binding site.

2.3.2.2. Pathway Methods

The change in free energy between two states of a system, for instance before and after binding, can be formally expressed as $-RTln\langle e^{-\Delta U/RT}\rangle$, where ΔU is the change in the potential energy between the two states. Here, the angle brackets indicate a Boltzmann average over a representative sample of conformations. For a binding reaction, ΔU represents the formation of interactions between the protein and the ligand, and the average can, in principle, be obtained by MD or Monte Carlo (MC) simulations. In reality, such a simulation is extremely difficult to converge. The free energy perturbation (FEP) method overcomes this drawback by breaking the change into N steps, each representing a small change δU of the interaction energy (perturbations). Then, a separate simulation is performed for each resulting energy function U_i to obtain the stepwise free energy changes $-RTln\langle e^{-\delta U/RT}\rangle_i$ associated with each step, where $\langle\ \ \rangle_i$ indicates a Boltzmann average with the energy function U_i. In this manner, the initial and

final states are connected by a continuous pathway of small steps for which relatively small free energy differences are computed.

Thermodynamic integration (TI) is another pathway approach. TI uses MD or MC to compute the first derivative of the free energy with respect to the distance along the path and afterwards estimates the total change in free energy through a numerical integration of the derivative along the path. Since the first derivative of the free energy is a force, TI effectively involves a work integral. Either free energy perturbation or thermodynamics integration can be applied to a given free energy calculation.

The first applications of these methods to binding processes aimed at computing the difference between the binding free energy of two similar ligands ($\Delta\Delta G^o$), based on a method called computational alchemy[22]. This involves using pathways to calculate the change in free energy when ligand A transformed into ligand B both in the binding site of the protein as well as in solution.

Pathway methods can also be employed to compute the standard binding free energy of a protein and a ligand. The Double Decoupling approach draws on related approaches[23] especially the Double Annihilation method[24]. It is based on a pathway technique to compute the work of gradually decoupling the ligand from the binding site and then effectively coupling it with an energy well or trap of the defined size in bulk solution. The work of allowing the ligand to escape from the trap into a standard volume can be computed analytically. Thus the simulations results are connected with the appropriate standard state. This approach has been further developed and applied to several protein-ligand systems.

Alternative approaches compute the free energy of binding via a pathway in which the ligand is gradually extracted from the binding site. In techniques of this sort, the free energy of binding is commonly derived from the potential of mean force extracted from a series of MD simulations.

2.4. Free Energy Differences From Simulations of Intermediate States

2.4.1. Exponential averaging

Let us assume that states i and j are defined by energy functions as a function of the particle coordinates $\vec{q}$, $U_i(\vec{q})$ and $U_j(\vec{q})$. Let ΔA_{ij} be the free energy difference between states i and j, defined as the logarithm of the ratio of the partition functions related to $U_i(\vec{q})$ and $U_j(\vec{q})$ ($Q_i$and Q_j). It has long been known that the free energy difference between two states can be computed by taking the exponential average of the potential energy differences [25].

$$\Delta A_{ij} = A_j(\vec{q}) - A_i(\vec{q}) = -\beta^{-1} \ln\left(\frac{Q_j}{Q_i}\right) \quad (2.27)$$

By adding and subtracting $e^{-\beta U_i(\vec{q})}$ from the integral in the partition function in the numerator, we get

$$\Delta A_{ij} = -\beta^{-1} \ln\left(\frac{\int e^{-\beta\left(U_j(\vec{q}) - U_i(\vec{q}) + U_i(\vec{q})\right)} d\vec{q}}{Q_i}\right) \quad (2.28)$$

$$\Delta A_{ij} = -\beta^{-1} \ln\left(\frac{\int e^{-\beta\left(U_j(\vec{q}) - U_i(\vec{q})\right)} e^{-\beta U_i(\vec{q})} d\vec{q}}{Q_i}\right) \quad (2.29)$$

which gives the final relationship of

$$\Delta U_{ij} = -\beta^{-1} \ln \langle e^{-\beta\left(U_j(\vec{q}) - U_i(\vec{q})\right)} \rangle_i \quad (2.30)$$

Even though exponential averaging is an exact solution and one of the simplest methods to understand, it is also one of the poorest methods in terms of efficiency. The exponential averaging method does not converge quickly with the number of samples.

2.4.2 Thermodynamic Integration

Starting with the statistical relation of the free energy

$$A = -\beta^{-1} lnQ \quad (2.31)$$

and taking the derivative with respect to λ

$$\frac{dA}{d\lambda} = -\beta^{-1} \frac{d}{d\lambda} ln \int e^{-\beta U(\lambda,\vec{q})} d\vec{q} = \langle \frac{dU(\lambda,\vec{q})}{d\lambda} \rangle_{\lambda} \quad (2.32)$$

one can obtain the final TI equation

$$\Delta A = \int_0^1 \langle \frac{dU(\lambda,\vec{q})}{d\lambda} \rangle_{\lambda} \, d\lambda \quad (2.33)$$

Because one can only carry out simulations at a number of intermediate values of λ, numerical integration schemes are required. All numerical integration schemes that are used in practice have the following form:

$$\Delta A \approx \sum_{k=1}^{K} w_k \langle \frac{dU(\lambda,\vec{q})}{d\lambda} \rangle_k \quad (2.34)$$

where w_k stand for the weights, which depend on the chosen numerical integration method[26, 27]. For instance, under the trapezoid rule even λ spacing weights are $w_1 = w_K = 1/2(K-1)$ and $w_{k\neq 1,K} = 1/(K-1)$.

2.4.3. Bennett Acceptance Ratio

The Bennett Acceptance Ratio (BAR) is one of the earliest methods for free energy estimation. BAR uses data from multiple states to give reliable free energy difference estimation. Both exponential averaging and thermodynamic integration need the ensemble average from a single state to estimate free energies. Even though TI requires the derivatives at state k, it does not need the configurations from any neighboring state. However, BAR requires configurational information from two states to estimate the free energy change. BAR is based on the principle that at the same configuration, $\vec{q}$, at two separate states, i and j, there exists a pathway connecting the two potentials, $U_i(\vec{q})$ and $U_j(\vec{q})$, and the difference $\Delta U_{ij}(\vec{q})$. In other words, the states i and j are defined by two different potentials acting on the same configuration. Owing to the states are in the same configuration, there is an exact relationship between the

distributions of the potential energy differences $\Delta U_{ij}(\vec{q})$ of the states sampled from i and $\Delta U_{ji}(\vec{q})$, what is the distribution of potential energy differences sampled from state j. Because it is an exact function of distributions, statistics can be applied to find the optimal way to use the information between two states, improving the free energy estimate.

Taking the properties of expectation values, we can write the free energy difference:

$$\Delta A_{ij} = -k_\beta T \ln\left(\frac{Q_j}{Q_i}\right) = -k_\beta T ln \frac{\langle \alpha(\vec{q}) e^{-\beta \Delta U_{ij}(\vec{q})} \rangle_j}{\langle \alpha(\vec{q}) e^{-\beta \Delta U_{ji}(\vec{q})} \rangle_i} \quad (2.35)$$

which is true for any $\alpha(\vec{q}) > 0$ for all $\vec{q}$. This is where Bennett started to drive the equation named after him[28]. Then he used variational calculus to select the value of $\alpha(\vec{q})$ minimizing the variance of the free energy.

$$\sum_{i=1}^{n_i} \frac{1}{1 + e^{\ln(n_i/n_j) + \beta \Delta U_{ij}(\vec{q}) - \beta \Delta A}} - \sum_{j=1}^{n_j} \frac{1}{1 + e^{\ln(n_j/n_i) + \beta \Delta U_{ji}(\vec{q}) - \beta \Delta A}} = 0 \quad (2.36)$$

which must be solved numerically. For the detailed derivation of the equation (2.36) see the original paper by Bennett[28]. Maximum likelihood is another approach to derive the same equation[29]. BAR is used to estimate free energies between many states; however, it can only do this in a pairwise manner. Thus, BAR requires information collected at state k and at its two neighboring states $k-1$ and $k+1$ for each configuration stored.

2.4.4. Potential of Mean Force Using Weighted Histogram Analysis Method

The potential of mean force (PMF) $W(\xi)$ along some coordinate ξ, was first introduced by Kirkwood[30], is a key concept in modern statistical mechanical theories of liquids and biomolecules. The PMF is defined based on the average distribution function $\langle p(\xi) \rangle$[31],

$$\mathcal{W}(\xi) = \mathcal{W}(\xi^*) - k_\beta T ln \left[\frac{\langle p(\xi) \rangle}{\langle p(\xi^*) \rangle}\right] \quad (2.37)$$

where ξ^* and $\mathcal{W}(\xi^*)$ are arbitrary constants. Using Boltzmann weighted

averages one can define the average distribution function along the coordinate ξ as follows

$$\langle p(\xi)\rangle = \frac{\int dR\delta\,(\xi'(R)-\xi)e^{-U(R)/k_\beta T}}{\int dR e^{-U(R)/k_\beta T}} \quad (2.38)$$

where $U(R)$ stands for the total energy of the system as a function of the coordinates R and $\xi'(R)$ is a function depending on a few degrees of freedom in the dynamical system. For instance, $\xi'(R)$ may be the distance between the centers of mass of two molecules.

It is generally not applicable to compute the PMF $\mathcal{W}(\xi)$ or the distribution function $\langle p(\xi)\rangle$, from a plain molecular dynamics simulation. The presence of large energy barriers along ξ may prevent an accurate sampling of the configurational space within the simulation time. For this reason, special techniques have been developed to calculate the PMF from molecular dynamics trajectories. The umbrella sampling technique by Torrie and Valleau[32] is one of those approaches. In this method the system of interest is simulated in the presence of a biasing window potential $w(\xi)$ to enhance the sampling in the neighborhood of a prescribed value of ξ. Thus, the biased simulations are performed using the potential energy $U(R)+w(\xi)$. In this manner, the biasing potential serves to confine variations of the coordinate ξ within a small interval around some chosen value and improve the configurational sampling in this particular region. For instance, a harmonic potential of the form $w(\xi)=\frac{1}{2}K(\xi-\xi_i)^2$ centered on successive values of ξ_i is a reasonable choice to produce the biased ensemble. To obtain the PMF along the region of interest of ξ, it is necessary to carry out a number of simulations each biasing the configurational sampling about a different value of the reaction coordinate ξ. Then, the results of the various windows are unbiased and recombined together to obtain the final estimate of $\mathcal{W}(\xi)$.

In umbrella simulations calculations, the last steps are very important. The biased distribution function obtained from ith-biased ensemble is as follows

$$\langle p(\xi)\rangle_{(i)} = e^{-w_i(\xi)/k_\beta T}\langle p(\xi)\rangle\langle e^{-w_i(\xi)/k_\beta T}\rangle^{-1} \quad (2.39)$$

The unbiased PMF from the ith window is

$$\mathcal{W}_i(\xi) = \mathcal{W}(\xi^*) - k_\beta T ln\left[\frac{\langle p(\xi)\rangle_{(i)}}{\langle p(\xi^*)\rangle}\right] - w_i(\xi) + F_i \quad (2.40)$$

where the undetermined constant F_i is defined as.

$$e^{-F_i/k_\beta T} = \langle e^{-w_i(\xi)/k_\beta T}\rangle \quad (2.41)$$

F_i represents the free energy associated with introducing the biasing potential. There have been a number of efforts for addressing the problem of unbiasing and recombining the information from umbrella sampling calculations.

For example, the Weighted Histogram Analysis Method (WHAM) by Kumar et al.[33] aims at using all information from umbrella sampling simulations. This method is a generalization of the histogram method that was developed by Ferrenberg and Swendsen[34]. In this method an optimal estimate of the unbiased distribution function is constructed as a weighted sum over the data extracted from all the simulations and then, the functional form of the weight factors that minimizes the statistical error is determined.

Let us consider an umbrella sampling calculation involving N_w biased window simulations. The WHAM equations express the optimal estimate for the unbiased distribution function as a ξ-dependent weighted sum over N_w individual unbiased distribution function $[\langle p(\xi)\rangle]_{(i)}^{unbiased}$.

$$\langle p(\xi)\rangle = \sum_{i=1}^{N_w}[\langle p(\xi)\rangle]_{(i)}^{unbiased}\left[\frac{n_i e^{-[w_i(\xi)-F_i]/k_\beta T}}{\sum_{j=1}^{N_w} n_j e^{-[w_i(\xi)-F_j]/k_\beta T}}\right] \quad (2.42)$$

In the above equation n_i is the number of independent data points used to construct the biased distribution function. Based on the previous equations one can write the individual unbiased distribution function as

$$[\langle p(\xi)\rangle]_{(i)}^{unbiased} = e^{+w_i(\xi)/k_\beta}\langle p(\xi)\rangle_{(i)}e^{-F_i/k_\beta T} \quad (2.43)$$

The WHAM equation can be rewritten as

$$\langle p(\xi)\rangle = \sum_{i=1}^{N_w} n_i \langle p(\xi)\rangle_{(i)} \left[\sum_{j=1}^{N_w} n_j e^{-[w_i(\xi)-F_j]/k_\beta T}\right]^{-1} \qquad (2.44)$$

The free energy constants F_i are determined from using the optimal estimate for the distribution function

$$e^{-F_i/k_\beta T} = \int d\xi e^{-w_i(\xi)/k_\beta T}\langle p(\xi)\rangle \qquad (2.45)$$

Since the distribution function itself depends on the set of constants $\{F_j\}$, the WHAM equations must be solved self-consistently. This is achieved through starting from an initial guess for the N_w free energy constants F_i, and an estimate for the unbiased distribution is obtained from equation 2.44. This estimate for $\langle p(\xi)\rangle$ is plugged in equation 2.45 to generate new estimates for the N_w free energy constants F_i and a new unbiased distribution is generated with equation 2.44. These steps are repeated until both equations are satisfied[31].

3.4.5. Multistate Bennett Acceptance Ratio

The estimator Multistate Bennett Acceptance Ratio (MBAR) is a direct extension to BAR as it enables assessing data from all states. MBAR reduces to BAR when only two states are considered and to WHAM in the limit that histogram bin widths are shrunk to zero. MBAR bases on the extended bridge sampling estimator[35].

Suppose we have N_i uncorrelated equilibrium samples from K thermodynamic states within the same ensemble, such as NPT, NVT, or μVT. We define the reduced potential function $u_i(x)$ for state i as follows

$$u_i(x) = \beta_i[U_i(x) + p_i V(x) + \mu_i^T n(x)] \quad (2.46)$$

where $x \in \Gamma$ indicates the configuration of the system within a configuration space with volume $V(x)$ (in the case of a constant pressure ensemble) and $n(x)$ the number of molecules of each of M components of the system (in the case of a grand-canonical ensemble). For each state i, β_i and $U_i(x)$ denote the inverse

temperature and the potential energy function, respectively, p_i stands for the external pressure and μ_i denotes the vector of the chemical potentials of the M state components.

Configurations $\{x_{in}\}_{n=1}^{N_i}$ from state i are sampled from the probability distribution

$$p_i(x) = c_i^{-1} q_i(x);\ c_i = \int_\Gamma dx\, q_i(x) \quad (2.47)$$

where $q(x)$ is nonnegative and represents an unnormalized density function, and c_i is the normalization constant (the partition function). In samples obtained from molecular dynamics, the unnormalized density is the Boltzmann's weight $q(x) = e^{u_i(x)}$.

The difference in dimensionless free energies

$$\Delta f_{ij} = f_j - f_i = ln \frac{c_j}{c_i} = ln \frac{\int_\Gamma dx\, q_j(x)}{\int_\Gamma dx\, q_i(x)} \quad (2.48)$$

where the f_i are related to the dimensional energies F_i by $f_i = \beta_i F_i$ and also the equilibrium expectations.

$$\langle A \rangle_i = \int_\Gamma dx\, p_i(x) A(x) = \frac{\int_\Gamma dx\, A(x) q_i}{\int_\Gamma dx\, q_i(x)} \quad (2.49)$$

One can compute these expectation values as ratios of the partition functions. In order to construct an estimator to compute these ratios we need the following relation

$$c_i \langle \alpha_{ij} q_j \rangle_i = \left[\int_\Gamma dx\, q_i(x) \right] . \frac{\int_\Gamma dx\, q_i(x) \alpha_{ij}(x) q_j(x)}{\int_\Gamma dx\, q_i(x)} \quad (2.50)$$

$$= \int_{\Gamma} dx\, q_i(x)\alpha_{ij}(x)q_j(x) \quad (2.51)$$

$$= \left[\int_{\Gamma} dx\, q_j(x)\right].\frac{\int_{\Gamma} dx\, q_j(x)\alpha_{ij}(x)q_i(x)}{\int_{\Gamma} dx\, q_j(x)} \quad (2.52)$$

$$= c_j\langle\alpha_{ij}q_i\rangle_j \quad (2.53)$$

which holds for any arbitrary choice of functions $\alpha_{ij}(x)$ for $c_i \neq 0$. Using the above relation, summing over the index j, and substituting the empirical estimator $N_i^{-1}\sum_{n=1}^{N_i} g(x_{ni})$ for the expectations $\langle g\rangle_i$, we obtain a set of K estimating equations

$$\sum_{j=1}^{K}\frac{\hat{c}_i}{N_i}\sum_{n=1}^{N_i}\alpha_{ij}q_j(x_{in}) = \sum_{j=1}^{K}\frac{\hat{c}_j}{N_i}\sum_{n=1}^{N_j}\alpha_{ij}q_i(x_{jn}) \quad (2.54)$$

for $i = 1,2,3,\ldots,K$, where the solution of the set of equations for the $\hat{c}_i$ yields an estimate of the c_i from the sampled data determined up to a multiplicative constant.

Equation 2.54 determines a set of asymptotically unbiased estimators that depend on the choice of functions $\alpha_{ij}(x)$, known as extended bridge sampling estimators in statistics[36]. By choosing the $\alpha_{ij}(x)$ as following

$$\alpha_{ij}(x) = \frac{N_j\hat{c}_j}{\sum_{k=1}^{K} N_k\hat{c}_k^{-1}\, q_k(x)} \quad (2.55)$$

we ensure that the resulting estimator is optimal, that it has the lowest variance for a large class of choice of $\alpha_{ij}(x)$ [36], and is guaranteed to have an unique solution up to a scalar multiplier[35].

When the configurations are obtained from Boltzmann statistics, where $q_i(x) \equiv e^{-u_i(x)}$, equations 2.54 and 2.55 yield the following estimating equations for the dimensionless free energies

$$\hat{f}_i = -ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{e^{-u_i(x_{jn})}}{\sum_{k=1}^{K} N_k e^{(\hat{f}_k - u_k(x_{jn}))}} \quad (2.56)$$

which must be solved in a self-consistent manner for the $\hat{f}_i$. Here we must emphasize that since the partition functions are only determined up to a scalar multiplier, the estimated free energies $\hat{f}_i$ can only be determined unequally up to an additive constant. Therefore, only calculations of the differences $\Delta\hat{f}_{ij} = \hat{f}_j - \hat{f}_i$ will be meaningful.

3. Energetics of Hydrophilic Protein-Protein Association and Role of water

The project described in this chapter has been published in the Journal of Chemical Theory and Computation (2014), 10, 3512-3524.

3.1. Summary

Hydrophilic protein-protein interfaces constitute a major part of all protein-protein interfaces and are thus of great importance. However, the quantitative characterization of their association is still an ongoing challenge and the driving force behind their association remains poorly characterized. In this chapter we addressed the association of hydrophilic proteins and the role of water by means of extensive molecular dynamics simulations in explicit water using three well studied protein complexes; Barnase - Barstar, Cytochrome *c* – Cytochrome *c* peroxidase, and the N-terminal domain of enzyme I – Histidine-containing Phosphocarrier. The one-dimensional free energy profiles obtained from umbrella sampling simulations are downhill or, in other words, barrierless. Using these one-dimensional free energy profiles, the computed standard free energies of binding are -12.7 ± 1.1 kcal/mol**,** -9.4 ± 0.7 kcal/mol and -8.4 ± 1.9 kcal/mol which are in reasonable to very good agreement with the experimental values of -19.6 kcal/mol, -8.8 kcal/mol and -7.8 kcal.mol. As expected, analysis of the confined water between the hydrophilic complex partners shows that the density and the orientational order parameter deviate noticeably from the bulk values, especially at close separations of the confining proteins.

3.2. Introduction

The capability of proteins to bind each other in a specific manner is essential for a wide variety of biological processes such as the cell cycle, cellular transport, immune response, apoptosis, DNA replication and transcription, RNA splicing and signal transduction. Even though many proteins perform their functions independently, a large part of all proteins interact with others for proper

biological activity. Based on large-scale proteomics studies, it has been estimated that about half of all cellular proteins are permanently or transiently involved in protein complexes where they form on average 6 - 8 interactions with other proteins[37]. Studies that analyzed the residue composition of protein-protein interfaces showed that protein-protein interfaces are enriched in both charged and polar residues rather than nonpolar residues[38, 39]. Yet, hydrophobic residues are found to be scattered over the entire interface where they form small hydrophobic patches within polar and charged residues[40].

Similar to the protein folding process, the decrease in the total Gibbs free energy (at constant temperature and pressure) of the protein-protein-solvent system upon binding is accompanied by opposing roles of entropy and enthalpy. For the protein folding process, the requirement for lowering the free energy while reducing the conformational space ensures that the energy landscape is funnel-shaped[2, 41]. Analogously, spontaneous protein-protein association also lowers the free energy of the full system while reducing the conformational, rotational and translational entropies of the binding partners[3]. Decades of experimental and theoretical efforts in this field have led to the development of paradigms, theoretical methods and computational tools. An established concept is the calculation of the standard free energy of binding from a one-dimensional potential of mean force (PMF). The potential of mean force is the work required to bring two particles to a particular relative separation *r*. Its gradient gives the average force, including direct and indirect contributions. PMF calculations provide a reliable method for determining the absolute free energies of binding of protein-ligand and protein-protein systems[42]. The commonly used method for this purpose is umbrella sampling[43] with the Weighted Histogram Analysis Method (WHAM)[33]. This method is able to connect the mechanistic details of the binding process with the underlying free energy surface. Application of these methods is relatively straightforward for small systems. However, it is an ongoing challenge for large systems such as protein-protein complexes.

Another important factor is the water solvent that plays a crucial role in governing the structure, stability, dynamics, and function of biomolecules. In many natural systems, water is confined in an environment where its free movement is restricted and its three-dimensional hydrogen-bonded network is

partially disrupted. The properties of the confined water are difficult to predict and may be considerably different from those of bulk water. For example, hydrophobic dewetting has been reported as a general mechanism for the association of hydrophobic surfaces[5]. Our understanding how hydrophilic surfaces assemble is less clear. There is some evidence e.g. by McLain *et al.*[44] that the association of hydrophilic surfaces results from the direct interactions between the binding partners. Based on neutron diffraction data, these authors showed that the association of some small peptides in aqueous solution is dominated by charge-charge interactions among the solutes. On the contrary, based on thermodynamic models, Ben-Naim argued that solvent-induced interactions play a dominant role for protein-protein association[45]. Using model systems, Berne and coworkers[46] addressed this issue in detail. In their study, plates characterized by large hydrophobic domains interacted via attractive solvent induced interactions. In contrast, a homogeneous distribution of hydrophobic and hydrophilic particles on the plates produced repulsive solvent induced interactions. Ahmad *et al.*[47, 48] extensively studied the process of protein-protein association by unbiased MD simulations in explicit solvent. When studying the spontaneous binding of a proline-rich peptide to an SH3 domain they found a clear dewetting transition upon binding of this hydrophobic interface[47]. With respect to hydrophilic interfaces, they observed that the nature of the water confined in the interfacial gap volume between the hydrophilic protein interfaces of the Barnase:Barstar pair deviated significantly from bulk behavior[48].

Taking three well-studied protein-protein complexes as model systems, we revisit here the issue of hydrophilic protein-protein association and the role of water using extensive molecular dynamics simulations. The studied systems are Barnase-Barstar (BN-BS), Cytochrome *c* – Cytochrome *c* peroxidase (CC-CCP), and the complex of the N-terminal domain of enzyme I with Histidine-containing Phosphocarrier (EIN-HPr). The three model complexes were selected due to the availability of a wealth of kinetic and structural data, due to the demonstrated importance of electrostatic interactions on their association, and because none of them exhibits large conformational changes upon complexation.

BN-BS is one of the best studied protein-protein complexes. Barnase is an bacterial RNAase that is lethal to the cell when expressed without its intracellular inhibitor Barstar[49]. The association of the BN-BS pair was addressed extensively in previous studies using e.g., Brownian dynamics (BD) and molecular dynamics (MD) simulations. Using BD simulations, Gabdoulline and Wade[50] computed association rates and their dependence on ionic strength and protein mutations for the BN-BS system. Spaar and Helms[51] characterized the long-range free energy landscape of the BN-BS encounter complexes using BD simulations. Using a transition-state theory like expression and atomic-detail modeling of proteins, Zhou and colleagues[52] calculated the electrostatic enhancement of the association rate of BN-BS and obtained results that correlated well with experiments. Recently Gumbart *et al.*[53] calculated the standard binding free energy of the BN-BS system from atomistic MD simulations in explicit solvent. Their newly proposed methodology relies upon PMF calculations where the proteins are restrained in the conformation, relative position and orientation of the bound state.

Both Cytochrome *c* (CC) and Cytochrome *c* peroxidase (CCP) are located in the intermembrane space of mitochondria. Utilizing two molecules of ferrous CC as specific electron source, CCP catalyzes the two-electron reduction of alkyl hydroperoxides[54]. Using BD simulations Northrup *et al.*[55] studied the diffusional association of the CC-CC pair and, Gabdoulline and Wade[56] investigated the factors that influence association rates of five different protein-protein complexes including the CC-CC pair as model systems. They found, for example, that the CC-CC association rate is fast enough to support a two-step electron transfer mechanism.

In bacterial cells, the phosphorylation and the translocation of sugars are coupled by the phosphoenolpyruvate (PEP):sugar phosphotransferase system (PTS) that consists of two cytosolic proteins, namely enzyme I and HPr, as well as of sugar specific components. The complex between EIN and HPr is a classical example of surface complementarity[57]. Using data from paramagnetic relaxation enhancement and replica exchange simulations, Hummer and coworkers[58] studied the transient encounter complexes in the EIN-HPr

association. They reported that, besides the specific complex, distinct nonspecific complexes exist as well that account for ~10% of relative population.

Here, we present the results from umbrella sampling simulations in explicit water and simulations of water localized between two proteins for the three systems mentioned above. The one-dimensional free energy profiles of protein-protein association were found to be downhill. Using these one-dimensional free energy profiles, the computed standard free energies of binding are in overall good agreement with the experimental values. Decomposition of the free energy of binding revealed that the direct non-bonded interactions between the complex partners favor the association whereas the solvent-induced interactions turned out to be repulsive. Moreover, the density and the orientational order parameter of confined water deviate noticeably from the bulk values at close separation of the confining proteins.

3.3. Materials and Methods

The coordinates for the bound protein-protein complexes were retrieved from the protein databank[59]: Barnase-Barstar (PDB ID: 1BRS[49]), Cytochrome *c* - cytochrome *c* peroxidase (PDB ID: 2PCC[54]) and the amino terminal domain of enzyme I and the Histidine-containing Phosphocarrier protein (PDB ID: 3EZB[57]).

3.3.1. Parameterization of Proteins

The titration states of titratable amino acids were assigned at physiological pH using the program PROPKA (http://nbcr-222.ucsd.edu/pdb2pqr_1.8/)[60]. All crystallographically resolved water molecules were kept and the placement of additional water molecules in internal protein cavities was tested using the program DOWSER [61] and keeping only those with DOWSER energy below -12 kcal/mol. All interactions were modeled by the Amber force field FF99SB-ILDN[62]. Short range nonbonded interactions were computed up to 1.2 nm distance. Long range electrostatic interactions were treated by the particle mesh Ewald (PME)[63] method. Dispersion correction was applied to energy and pressure. Periodic boundary conditions were applied in all directions. Water molecules were modeled by the TIP3P[64] potential that is typically used

together with the AMBER force field. All simulations were carried out using the GROMACS package, version 4.5.4.[65].

3.3.2. Heme Center Parameterization

The active sites of Cytochrome *c* and Cytochrome *c* peroxidase contain a heme group each. In Cytochrome *c,* the heme is covalently bonded to the polypeptide chain and the iron atom in both oxidized and reduced states is six-coordinate low spin[66]. The central heme iron is ligated to both a histidine and methionine. Ccp, in its resting state, involves a noncovalent heme with a five-coordinate high-spin iron[66]. Here, the central iron atom is coordinated by a histidine residue at the fifth position. In our study we considered Cytochrome *c* and Cytochrome *c* peroxidase in their resting states where the iron atoms are Fe(II) and Fe(III), respectively.

Even though the coordinate set 2PCC (PDB ID) was used to start the MD simulations, we retrieved coordinates for the heme group and its coordinating residues from crystal structures determined at higher resolution as starting positions for the quantum mechanical calculations. The PDB entries 1YYC (1.23 Å) and 1ZBY (1.20 Å) were used for Cytochrome *c* and Cytochrome *c* peroxidase, respectively. The amino acids bonded to the heme centers were truncated at their β-carbons and hydrogens were added. All QM calculations were performed using Gaussian03[67]. For derivation of partial atomic charges, we followed the standardized protocol commonly used in combination with the original AMBER94 force field with some minor changes. The geometry was optimized at B3LYP level using the basis set 6-31G* in two consecutive stages. First, we optimized the geometry of the heme group alone. Then the coordinating residues were added to the resulting heme configuration in their conformations observed in the crystallographic structure and the geometry of the full system was further optimized without any restraints. Using the optimized geometry, we obtained the molecular electrostatic potential from the HF/6-31G* electron density of the heme centers. Restricted ESP (RESP) charges[68] were obtained using the RESP program under Amber Tools in two steps. In the first stage, charge equivalency on chemically equivalent heavy atoms was imposed and the total charges on the methyl groups that were generated by adding hydrogen atoms to the $C\beta$ atoms of

the heme coordinating amino acids were set to zero. In the second stage, the charges of chemically equivalent hydrogen atoms were equated, keeping the constraint on methyl groups and the charges on heavy atoms constant. The remaining excess charge obtained when linking the amino acid $C\beta$ to the rest was equally distributed over all atoms included in the parameterization, ensuring that the overall charge on the heme and the newly defined amino acids is integral.

The missing force constants for bonds, angles and dihedrals of the heme groups were taken from heme parameters in the AMBER parameter database, see http://www.pharmacy.manchester.ac.uk/bryce/amber/ and Shahrokh *et al*.[69] The derived charges and the used force constants for the heme centers are provided in the Supporting Information.

3.3.3. Molecular Dynamics Simulations

In this study, we conducted two different sets of molecular dynamics simulations for the three protein-protein complexes in explicit solvent. First, we combined umbrella sampling and the weighted histogram analysis method to characterize the one dimensional binding free energy surface of the three systems. In order to generate equilibrated initial structures for the simulations, each system was placed in a cubic box of TIP3P water. To mimic physiological conditions 100 mM NaCl was added, including neutralizing counterions. This resulted in 50 Na^+ and 46 Cl^- for the BN-BS system, 83 Na^+ and 82 Cl^- for the CC-CC system and 107 Na^+ and 86 Cl^- for the EIN-HPr system. Following an initial energy minimization of 1000 steps of steepest descent, each system was equilibrated in two steps where the heavy atoms of the proteins were restrained using a force constant of 1000 kJ mol^{-1} nm^{-2}. The first step involved 100 ps of MD in the NVT ensemble, maintaining the temperature at 310 K. Protein and nonprotein atoms were coupled separately to temperature baths using Berendsen's weak coupling algorithm[70] with a coupling time of 0.1 ps. All bonds were constraint using LINCS algorithm[71]. Subsequently, 100 ps of NPT equilibration were performed, keeping the pressure at 1 bar also using Berendsen's weak coupling method[70] with a time constant of 1 ps. During data collection, the Nose-Hoover thermostat[72, 73] was combined with the Parrrinello-Rahman barostat[74] to regulate temperature and pressure,

respectively. Equilibration was completed by 20 ns of conventional MD simulation in the absence of any restraints. For integrating Newton's equations of motion, a leap-frog algorithm with a time step of 2 fs was used.

The final coordinates at the end of these trajectories were used as starting configurations for umbrella sampling simulations. The protein coordinates were rotated in order to align the line connecting the centers of mass of the two complex partners with the z-axis. Then, the two proteins were placed in a rectangular box with dimensions sufficient to satisfy the minimum image convention even at the largest separation distance of 3 nm. The solvation box sizes were 6.7 nm x 7.7 nm x 14.7 nm, 9.0 nm x 9.4 nm x 16.0 nm and 11.0 nm x 7.8 nm x 16.3 nm for BN-BS, CC-CC and EIN-HPr respectively. The initial configurations were generated by translating one of the protein partners along the z-axis up to 3 nm distance between the surfaces of the two proteins, while keeping the other one fixed. The disassociation path between the bound and the 3.0-nm separated states was divided into 0.1 nm intervals up to 1 nm of separation and 0.2 nm intervals between 1 nm and 3 nm. This resulted in 21 windows. For each window an umbrella sampling run of 10 ns length was performed with a force constant of 1000 kJ/mol.nm^2 in general. The force constant was increased to 2000 kJ/mol.nm^2 in cases when the protein centers in the windows deviated strongly from the initial configurations and therefore caused sampling problems. For the BN-BS system each window was extended to 40 ns to check whether this led to a better agreement with the experimental binding free energy. The first 0.5 ns of all windows were considered as equilibration and excluded during analysis. Different time intervals were utilized for construction of the PMF curves.

For WHAM analysis, the g_wham[75] utility of GROMACS 4.5.4 was used with default options, except for the convergence tolerance that was set to 10^{-9}. The histograms were carefully analyzed to ensure sufficient overlapping. New windows were added when the overlaps between the histograms were not sufficient and existing windows were deleted in those cases to avoid redundancy. The histograms of the final umbrella windows are displayed in Figure 3.1.

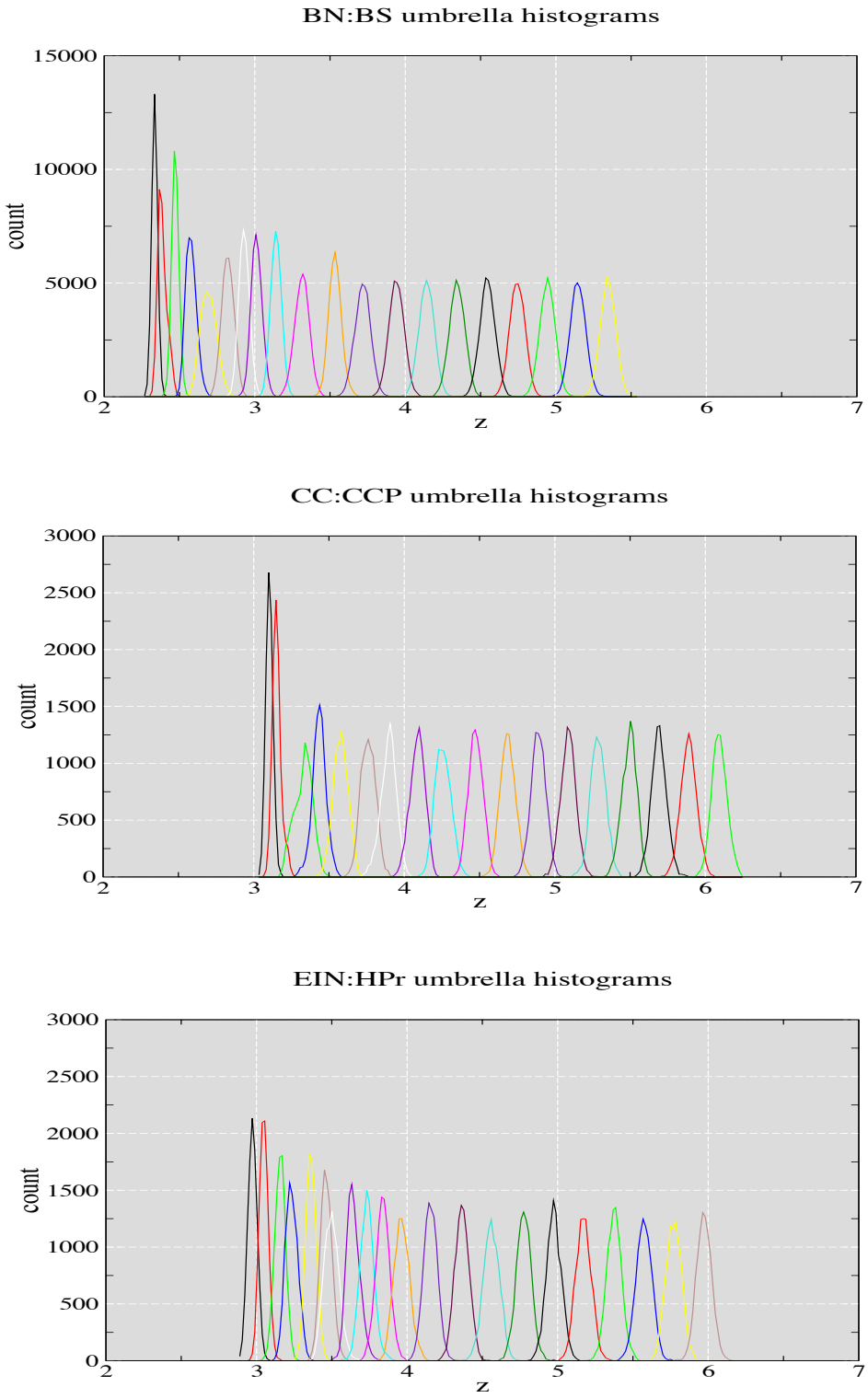

Figure 3.1. Histograms of the configurations within the umbrella sampling windows for BN-BS (upper panel), CC-CC (middle panel) and EIN-HPr (lower panel). z is the reaction coordinate coincident with the COM distance between the protein partners.

3.3.4. Standard Free Energy of Binding from PMF

In order to calculate the standard free energy of binding, ΔG^o, we followed the strategy presented by Henchman and coworkers[76]. They assumed that a PMF is sampled along a 1D reaction coordinate, whereas concurrently the orthogonal translational movement is restricted by a harmonic confinement potential. No restraints are applied to the orientations. The standard free energy of binding is computed as the sum of three terms.

$$\Delta G^o = \Delta G_{PMF} + \Delta G_V + \Delta G_R \quad (3.1)$$

where ΔG_{PMF} stands for the binding free energy change obtained as the difference between the bound and unbound states retrieved from the PMF, ΔG_V stands for the free energy change from the unbound volume to the standard state volume and ΔG_R accounts for the change in free energy associated with the introduction of translational confinement restraints.

ΔG_{PMF} and ΔG_v are computed as

$$\Delta G_{PMF} = -RTln\left(\frac{Q_b}{Q_u}\right) \quad (3.2)$$

and

$$\Delta G_v = -RTln\left(\frac{V_u}{V_o}\right) \quad (3.3)$$

where, R is the ideal gas constant, T is the absolute temperature, V_u is the unbound volume, and V_o is the standard state volume. Q_b and Q_u are the partition functions for the bound and unbound regions, respectively. Their ratio is computed by the following equation:

$$\frac{Q_b}{Q_u} = \frac{l_b}{l_u} e^{\left(\frac{-\Delta W}{RT}\right)} \quad (3.4)$$

Here, the PMF depth, ΔW, is defined as the lowest point minus the exponential average over the entire unbound region of the PMF. $W(z)$ is the PMF as a function of z and defined to be zero at its lowest point when the proteins are bound.

$$\Delta W = RTln\left[\frac{\int_{unbound} e^{\frac{-W(z)}{RT}} dz}{\int_{unbound} dz}\right] \quad (3.5)$$

l_b and l_u are the configurational integral of the PMF and are given by the following equations:

$$l_b = \int_{bound} e^{\frac{-W(z)}{RT}} dz \quad (3.6)$$

and

$$l_u = \int_{unbound} dz \quad (3.7)$$

The unbound volume V_u is the area explored in the xy plane times the distance sampled along the z axis, l_u, by the protein and computed as follows:

$$V_u = l_u \frac{2\pi RT}{k_{xy}} \quad (3.8)$$

where k_{xy} is the force constant of the applied harmonic restraint potential along x and y directions.

The free energy term ΔG_R to remove the orthogonal restraints in the bound state is computed from additional 10 ns or 40 ns long umbrella windows without orthogonal restraints, which cover the bound region along the reaction coordinate, using the following equation

$$\Delta G_R = RTln\langle e^{\frac{-k_{xy}(\Delta x^2+\Delta y^2)}{2RT}}\rangle_{k_{xy}=0} \quad (3.9)$$

Here Δx and Δy are the observed displacements relative to the minimum in the restrained simulations.

In this study we chose the cutoff between the bound and unbound regions to be the value of z where the PMF becomes constant within statistical noise.

3.3.5. Rotational Entropy Calculation

We calculated the rotational entropy based on the distribution of the orientations of BS with respect to BN, of CC with respect to CCP and of HPr with respect to EIN. For calculating the rotational entropy we used the umbrella sampling trajectories after removing the rotational motion of one of the complex partners. First of all, the reaction coordinate was divided in equal bins of 0.1 nm length (distance bins). Then the snapshots were assigned to the distance bins according to the protein-protein COM distance. Afterwards, for each distance bin we calculated the entropy value as explained below.

The rotation matrices needed for calculating protein orientations were obtained by the g_rotmat routine of GROMACS that is based on least squares fitting. Utilizing the resulting matrices the three Euler angles (ϕ, θ, ψ) were computed. The sampled distribution of the Euler angles was used to compute the entropy as follows[77]:

$$S^{rot} = -R\int_{\phi=0}^{360} p(\phi)lnp(\phi)d\phi - R\int_{\theta=0}^{180} p(\theta)lnp(\theta)sin\theta d\theta - R\int_{\psi=o}^{360} p(\psi)lnp(\psi)d\psi \quad (3.10)$$

In order to define the states for the entropy calculation, the range for each Euler angle was equally divided in angular bins of 3.6° bin size. For every angular bin we counted the observed frequency, *p*. Summing up the contribution of each state according to the above formula gave the rotational entropy. Instead of the absolute rotational entropy we reported the change in rotational entropy, ΔS^{rot}, relative to the uniform distribution of the three Euler angles what corresponds to the ideal freely rotating case.

3.3.6. Entropy of Binding

The entropy loss of one of the protein partners upon association was also estimated using the SF (system-frame) method introduced by Irudayam and Henchman[78] for protein ligand systems. Here we give a brief description of the method. For the detailed derivations please see reference [78]. Even though in our systems both binding partners are proteins, we refer to the smaller binding partner (Barstar, Cytochrome *c* and Histidine-containing phosphocarrier) as ligand. We ignored the change in internal entropy and, thus, assume that the ligand only loses translational and rotational entropy upon complexation. In the unbound state, the translational entropy of the ligand, $S^{o,tr}_{L(aq)}$ is the sum of vibrational and cratic entropies and is given by

$$S^{o,tr}_{L(aq)} = Rln\left(\frac{1}{x^{o}_{L(aq)}}\prod_{i=1}^{3}\frac{2k_{\beta}Te}{F^{i}_{L(aq)}\Lambda_{L}}\right) \quad (3.11)$$

where $x^{o}_{L(aq)}$ is the mole fraction of L, e is the natural logarithm base, $F^{i}_{L(aq)}$ is half of the average force magnitudes along the principal axes of the ligand and Λ_L is the translational thermal de Broglie wavelength. The rotational entropy of the

ligand in the unbound state consists of orientational and librational terms and is given by

$$S_{L(aq)}^{rot} = Rln\left(\frac{8\pi^2}{\sigma_L V_{w(l)}}\prod_{i=1}^{3}\frac{2k_\beta Te}{r_{L(aq)}^i \tau_{L(aq)}^i \Lambda_L^i}\right) \quad (3.12)$$

In the above equation, σ_L is the symmetry number, $V_{w(l)}$ is the volume of a single water molecule, r_L^i is the radius of the ligand protein along the ith principal axis, τ_L^i is half of the torque magnitude about the ith principal axis of the ligand and Λ_L^i represents the rotational thermal de Broglie wavelengths.

In the bound state the ligand was assumed to have no cratic and orientational entropy. Hence, the translational entropy of the bound ligand, $S_{(comp,aq)}^{tr}$, is only vibrational

$$S_{(comp,aq)}^{tr} = Rln\left(\prod_{i=1}^{3}\frac{2k_\beta Te}{F_{L(comp,aq)}^i \Lambda_L}\right) \quad (3.13)$$

and the rotational entropy, $S_{(comp,aq)}^{rot}$, is solely librational

$S_{(comp,aq)}^{rot} = Rln\left(\prod_{i=1}^{3}\frac{2k_\beta Te}{\tau_{(comp,aq)}^i \Lambda_L^i}\right)$Taking the difference of the bound and free ligand entropies and assuming that the thermal rotational de Broglie wavelengths do not change between solution and the complex, the final equations for translational and rotational entropy changes upon complexation are as follows

$$\Delta S_L^{tr} = Rln\left(\frac{V_{w(l)}}{V^o}\prod_{i=1}^{3}\frac{F_{L(aq)}^i}{F_{L(comp,aq)}^i}\right) \quad (3.14)$$

$$\Delta S_L^{rot} = Rln\left(\frac{\sigma_L V_{w(l)}}{8\pi^2}\prod_{i=1}^{3}\frac{\tau_{(aq)}^i}{r_{L(aq)}^i \tau_{(comp,aq)}^i}\right) \quad (3.15)$$

The average force and torque magnitudes were extracted from the first and last windows of umbrella sampling simulations corresponding to the bound and unbound states, respectively.

3.3.7. Properties of Interfacial Water

Subsequently, we performed a second set of MD simulations for the three protein-protein systems to characterize the density and the tetrahedral order parameter of the interfacial water localized between the two protein interfaces.

These simulations were performed under the same conditions used in the umbrella sampling simulations. Additionally, harmonic position restraints were applied to the backbone atoms of the proteins to maintain their interfacial distance and a fixed relative orientation. The starting configurations were obtained in the same way as previously mentioned for the umbrella sampling simulations. The interfacial distance was varied between 0.35 and 5.0 nm. Each simulation was 20 ns long. The snapshots for analysis were collected every 0.25 ps. Analyses were carried out using the snapshots from the last 15 ns. A cubic box of pure water was also simulated in order to compare the properties of interfacial water to the bulk values.

3.3.8. Interfacial Gap Definition

The interface residues on the protein surfaces were retrieved from the ABC^2 database.[79] There, protein interfaces contain those residues exhibiting a certain change in their solvent accessible surface area (SASA) when comparing the values of the unbound state to those of the complexed forms. The interfacial gap was defined by a rectangular box, which spans the center of geometry of both interfaces along the z-axis (the length). The width and height of the box were calculated by considering minimum and maximum coordinate values of the interfaces along x- and y-axes. The values, which gave the smaller dimensions, were taken for defining the interfacial gap. Those water molecules, having oxygen positions inside the interfacial gap, were considered as interfacial water molecules. The volume of the interfacial water gap was calculated by subtracting the overlapping protein volume from the volume of the aforementioned rectangular box. The overlapping protein volume was calculated from the protein mass inside the rectangular box and assuming an average value of the protein density. For the protein density we used the generally accepted value of 1.35 g/cm^3 that was deduced from hydrodynamic[80, 81] and adiabatic compressibility[82] experiments. This procedure was applied to all frames along a trajectory and the quantities were averaged subsequently.

3.3.9. Tetrahedral Order Parameter

Water molecules have a general propensity for tetrahedral coordination, owing to their hydrogen-bonding network. The tetrahedral order parameter is a three body order parameter that measures the degree to which the nearest-neighbor molecules are tetrahedrally coordinated with respect to a given molecule[83, 84]. The tetrahedral order parameter is defined as

$$q = 1 - \frac{3}{8}\sum_{i=1}^{3}\sum_{j=i+1}^{4}\left(cos\psi_{ikj} + \frac{1}{3}\right)^2 \quad (3.16)$$

where ψ_{ikj} is the angle formed by the lines connecting the oxygen atoms of a given water molecule *k* and those of its nearest neighbors *i* and *j*. We computed the orientational order parameter in two different ways; first, we only considered water molecules as potential neighbors, second, in addition to the water molecules, we also took into account the nearby protein oxygen and nitrogen atoms within a cutoff 3.5 Å as potential neighbors. In both cases, we considered the four nearest neighbors of interfacial water molecule *k,* based on their Euclidian distances and calculated the orientational order parameter using the formula above. These calculations were performed for each interfacial water molecule *k* and then averaged.

3.4. Results and Discussion

3.4.1. Global and Interface Properties of the Systems of Interest

In this work we studied the association/dissociation of three well studied protein-protein complexes; Barnase-Barstar (BN-BS), Cytochrome *c* – Cytochrome *c* peroxidase (CC-CC) and the complex of the N-terminal domain of enzyme I with Histidine-containing Phosphocarrier (EIN-HPr). Figure 3.2 shows cartoon and electrostatic surface representations of the systems studied. Visual inspection easily reveals the favorable electrostatic complementarity between the protein partners, especially of the BN-BS and CC-CC complexes.

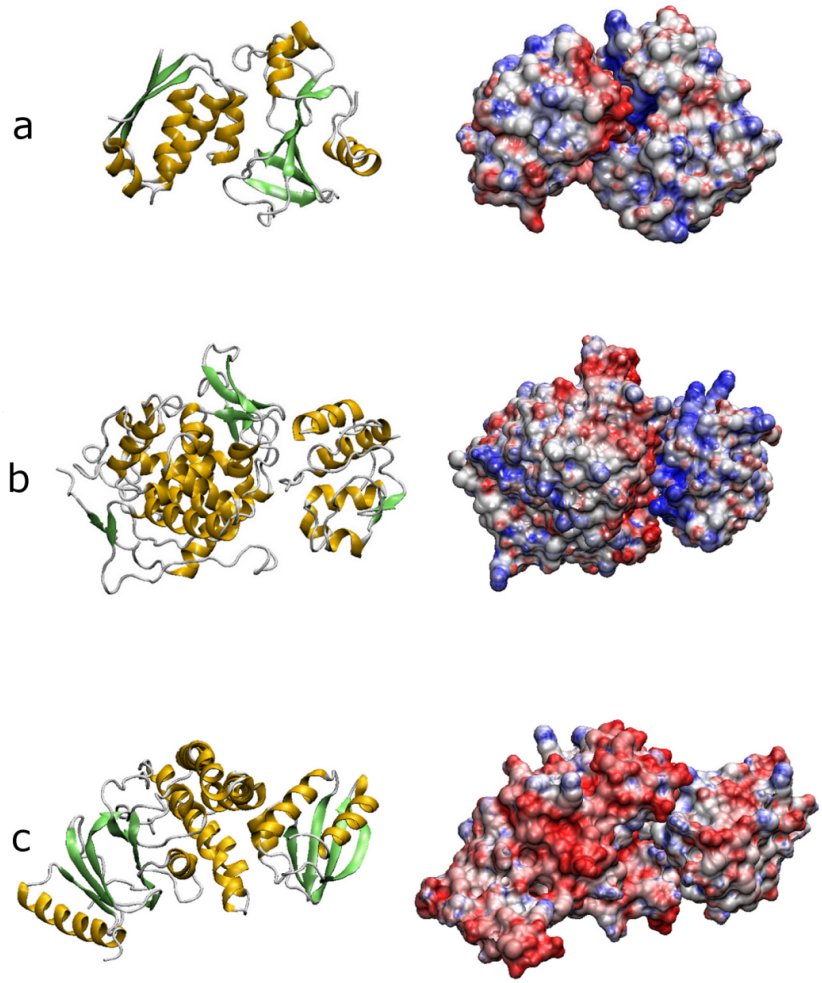

Figure 3.2. Cartoon (left panel) and electrostatic potential surface representation (right panel) of the studied complexes. a) Barnase: Barstar. b) Cytochrome c: Cytochrome c peroxidase c) N-terminal domain of enzyme I: Histidine-containing Phosphocarrier. The electrostatic potential was computed utilizing the Adaptive Poisson-Boltzmann Solver (APBS)[85] setting the dielectric constant to 78.5 and 2.0 for the solvent and the proteins, respectively. The electrostatic potential was mapped to the protein surface using the VMD software.[86] The color scale data ranges from -7kBT (red) to +7kBT (blue).

Table 3.1 summarizes general and interface properties of the three model systems. As seen in Table 3.1, the proteins and, more importantly, the interfaces carry non-zero electrostatic net charges at pH=7. The charge values given in Table 3.1 were computed after assignment of the titration states of amino acids by PROPKA[60] and may thus differ from the standard charge values reported in the literature. In all systems the interfaces are oppositely charged. In the BN-BS and CC-CC complexes, the proteins also have an opposite overall charge whereas in the EIN-HPr complex both proteins carry a negative overall charge. Another noteworthy property is the area of the binding interfaces. The EIN-HPr system

has the largest binding area which is commonly a characteristic of permanently assembled proteins[87]. Here, however, it is the system with the weakest binding constant. This may result from the overall negative charges of the complex partners. Among the known protein complexes, BN-BS is one of the tightest complexes with 14 hydrogen bonds and 12 salt bridges formed across the interfaces, bearing a K_d value of 1.3 x 10^{-14}. The CC-CC complex shows perfect electrostatic complementarity[56] but has the smallest binding interface of all three systems. The relatively weak stability of the bound complex may be connected to the transient nature of the electron transfer step taking place between this interaction pair.

Table 3.1. Some global and interface properties of the three protein-protein complexes. The interface area and number of interface residues were retrieved from the ABC[2] database[79]. The number of H-bonds and salt-bridges across the interfaces were taken from the PDBe Pisa database (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html).

	BN-BS	CC:CYP	EIN-HPr
Number of amino acids in protein I	110	108	249
Number of amino acids in protein II	89	296	85
Area of binding interface (Å^2)	778	570	1002
Number of interface residues in protein I	16	13	33
Number of interface residues in protein II	14	10	24
Number of H-bonds across interface	14	4	6
Number of salt-bridges across interface	12	2	5
Total charge of protein I [e]	+2	-7	-19
Total charge of protein II [e]	-6	+6	-2
Total charge of interface I [e]	+3	+5	-5
Total charge of interface II [e]	-4	-2	+4
Binding constant	1.3 x 10^{-14} M^{-1} [a]	6 × 10^{-7} M^{-1} [b]	3.1 × 10^{-6} M^{-1} [c]

a) The disassociation constant was retrieved from reference[88]. b) The disassociation constant was taken from reference[89]. c) The disassociation constant was converted from the association constant reported in reference[90].

3.4.2. One Dimensional Free Energy Surface of Protein-Protein Association

In this study we combined umbrella sampling with the WHAM method to obtain the PMF curve for protein-protein disassociation. Here, the reaction coordinate corresponds to the z-axis, coincident with the distance between the centers of mass (COM) of the complex partners. We partitioned the trajectory into pieces and computed the PMF and subsequently the standard free energy of binding. Figure 3.3 shows PMF curves for the three systems obtained using different time intervals of the full-length simulation windows. As seen in the figure, all one-dimensional free energy surfaces of association are downhill or in other words barrierless. For BN-BS, this was reported before[53, 91, 92]. Even though the maximum value of the PMF varies among different time intervals, the PMF curves behave similar for the three systems. Among the systems, the sharpest PMF curve belongs to the BN-BS complex. Here the PMF becomes flat at about 1.4 nm of separation what corresponds to 3.6 nm along the reaction coordinate. For the CC-CC and EIN-HPr systems the PMF curves flatten beyond 1.5 nm of separation and coincide with the data points at 4.6 nm and 4.5 nm on the reaction coordinate, respectively. The dashed line parallel to the y-axis in Figure 3.3 indicates this critical separation for each system.

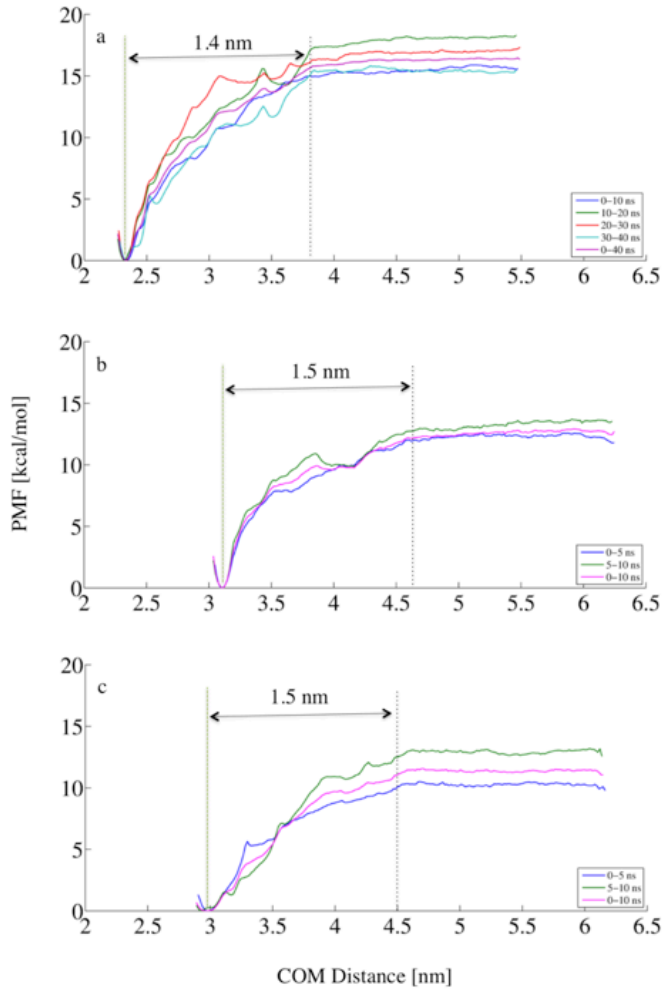

Figure 3.3. Potential of mean force calculated from different time intervals of the umbrella sampling simulations. a) Barnase: Barstar. b) Cytochrome *c*: Cytochrome *c* peroxidase c) N-terminal domain of enzyme I: Histidine-containing Phosphocarrier. The black dashed line that is parallel to the y-axis represents the cutoff that separates the bound region from the unbound region. The grey dashed line left of it marks the position of the bound state.

To obtain insight into the energetic contributions that lead to these PMF profiles, we analyzed the components of the non-bonded interaction terms (see Figure 3.4). The separation distances where the PMFs start to flatten coincide with the distance where the direct Lennard Jones (LJ) interactions between the proteins almost vanish along the reaction coordinate. In contrast, the direct electrostatic interactions are still very strong at these separations (see Figure 3.4). Interestingly, these separations are also where the rotational entropy of the proteins starts to converge to a constant value (see Figure 3.5).

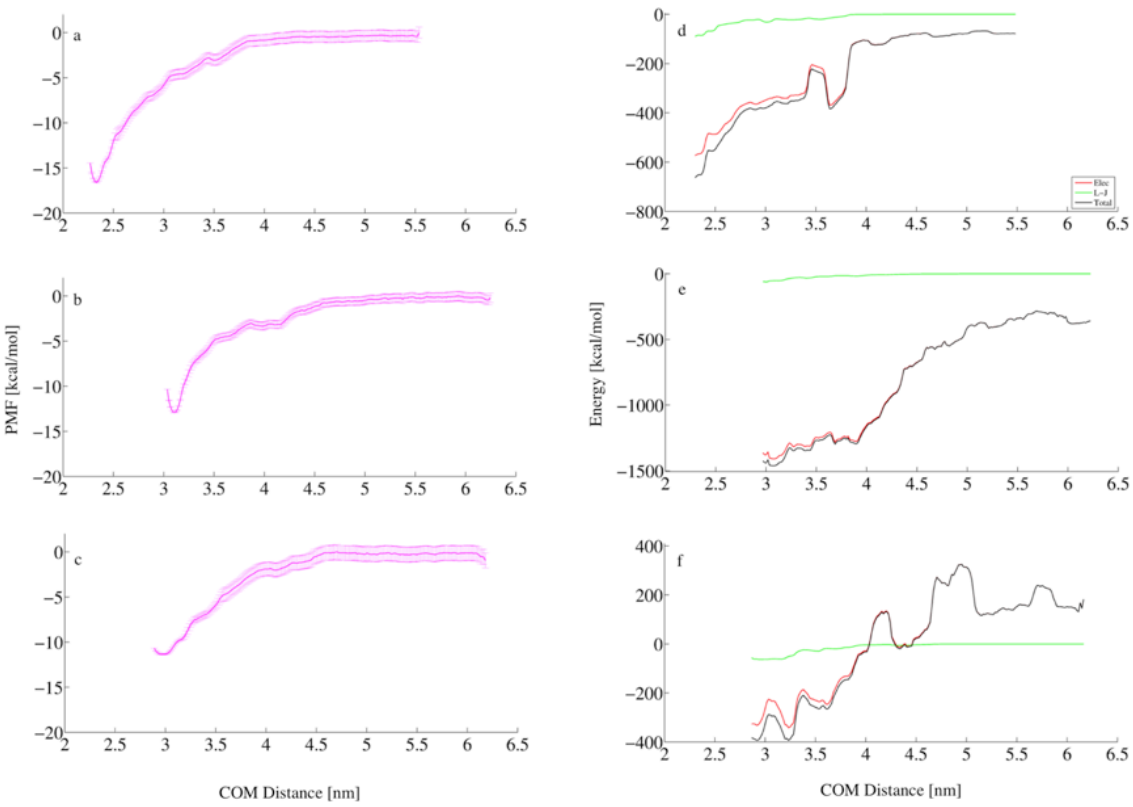

Figure 3.4. The potential of mean force (PMF) (left panel) and the direct interactions between the two proteins (right panel) calculated from full-length umbrella sampling simulation windows along the COM distance. a, d) Barnase: Barstar. b, e) Cytochrome *c*: Cytochrome *c* peroxidase. c, f) N-terminal domain of enzyme I: Histidine-containing Phosphocarrier. The error analyses for PMF values were performed using a bootstrap method introduced previously[75].

We assessed the quality of the PMF values obtained from different time intervals of the trajectory by inspecting the quality of the sampling histograms and made sure that there is sufficient overlapping between two consecutive windows. Not surprisingly, longer sampling times led to smoother histograms and better overlap. Assured by this analysis, we decided to use the full trajectories for further analysis.

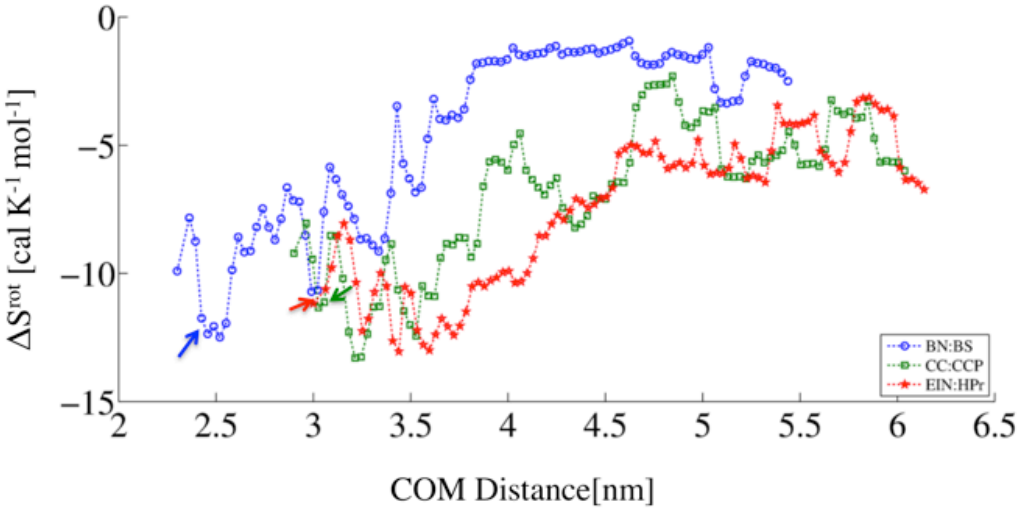

Figure 3.5. Rotational entropy change along the COM distance; Barnase: Barstar (blue). Cytochrome *c*: Cytochrome *c* peroxidase (green) and N-terminal domain of enzyme I: Histidine-containing Phosphocarrier (red). The rotational entropy is calculated based on the distribution of the orientations of BS with respect to BN, of CC with respect to CCP and of HPr with respect to EIN. The change in rotational entropy, ΔS^{rot}, is defined relative to the uniform distribution of the three Euler angles. Arrows mark the positions of the bound complexes.

3.4.3. Standard Free Energy of Binding

The relation between the equilibrium constant of a chemical reaction and the associated free energy is well established. When the reaction involves a change in the number of components, one has to relate the obtained results to a standard state, see reference [7]. Based on the PMF values given in Figure 3.5, we calculated the standard free energies of binding, as explained in the Methods section. The results are provided in Table 3.2 together with the experimental values for comparison.

The cutoff between bound and unbound regions was set to the value of the reaction coordinate where the PMF becomes constant within some error interval. In Figure 3.3, the cutoff distances are marked as black dashed lines parallel to the y-axis. Although the definition of the cutoff is arbitrary, the lowest values of the PMF in the binding site contribute the most to the integrals and make the calculation of ΔW insensitive to the cutoff[76]. Moreover, the rotational entropy values beyond these cutoffs do not change considerably meaning that the proteins are effectively in bulk (see Figure 3.5). This is more pronounced for the BN-BS system due to the better convergence during windows of 40 ns in length.

Table 3.2. Standard free energies of binding (ΔG^o) for the BN-BS, CC-CC and EIN-HPr complexes. The experimental values ΔG^o_{exp} are based on the binding constants given in Table 3.1. The calculated values; ΔW, ΔG_{PMF} ΔG_v and ΔG_R stand for the PMF depth, the free energy change of binding between the bound and unbound section of the PMF, the free energy change from the unbound volume to the standard state volume and the free energy change to remove the orthogonal restraints, respectively.

BN-BS						
Time interval	$\Delta W(\frac{kcal}{mol})$	$\Delta G_{PMF}(\frac{kcal}{mol})$	$\Delta G_v(\frac{kcal}{mol})$	$\Delta G_R(\frac{kcal}{mol})$	$\Delta G^o(\frac{kcal}{mol})$	$\Delta G^o_{exp}(\frac{kcal}{mol})$
0- 10 ns	-12.9	-13.5	2.5	-0.8	-11.8	-19.6[a]
10-20 ns	-15.3	-15.8	2.5	-0.9	-14.2	
20-30 ns	-14.2	-14.7	2.5	-0.8	-13.0	
30-40 ns	-12.9	-13.4	2.5	-0.9	-11.8	
-12.7 ± 1.1[d]						
0-40 ns	-13.7	-14.2	2.5	-0.9	-12.6	
CC-CC						
0-5 ns	-9.7	-10.3	2.5	-1.1	-8.9	-8.8[b]
5-10 ns	-10.6	-11.2	2.5	-1.2	-9.9	
-9.4 ± 0.7[d]						
0-10 ns	-10.2	-10.6	2.5	-1.2	-9.3	
EIN-HPr						
0-5 ns	-7.7	-8.6	2.5	-0.9	-7.0	-7.8[c]
5-10 ns	-10.3	-11.3	2.5	-0.9	-9.8	
-8.4 ± 1.9 [d]						
0-10 ns	-8.8	-9.9	2.5	-0.9	-8.3	

a-c) Converted from the disassociation constants reported in Table 3.1, using the simulation temperature (310 K). d) The mean value and the standard deviation of the standard free energies of binding ($\boldsymbol{\Delta G^o}$) from 10-ns (BN-BS) and 5-ns (CC-CC and EIN-HPr) time intervals.

The standard free energy of binding for the different time intervals varies between -11.8 - 14.2 kcal/mol, -8.9 -9.9 kcal/mol and -7.0 - 9.8 kcal/mol for BN-BS, CC-CC and EIN-HPr, respectively. These give average values of -12.7 ± 1.1, -9.4 ± 0.7 and -8.4 ± 1.9 kcal/mol, which are very close to the free energy of binding computed from the whole trajectory (-12.6 kcal/mol, -9.3 kcal/mol and -8.3 kcal/mol). Our estimates of the standard error of the mean standard free energy are based on the four 10-ns long trajectory parts for BN-BS and 5-ns long trajectory parts for CC-CC and EIN-HPr. We carefully searched the literature for

the best matching experimental conditions to our simulations. The calculated standard free energies of binding are in good agreement with these experimental values, except for BN-BS. For the CC-CC and EIN-HPr systems, the computed standard free energies of binding from 10-ns long trajectories, -9.3 kcal/mol and -8.3 kcal/mol, match very closely the experimental values of -8.8 kcal/mol and -7.8 kcal/mol, respectively. However for the BN-BS system the value computed from 10-ns long trajectories, -11.8 kcal/mol, is 7.8 kcal/mol less favorable than the experimental value of -19.6 kcal/mol. Extension of the simulations to 40 ns did not bring the computed value much closer to the experimental correspondent. The largest computed value is -14.2 kcal/mol, which is still 5.4 kcal/mol higher than the experimental value (see Table 3.2). It is well known that the computed standard free energy of binding based on the method we followed here is dependent on the applied orthogonal force constant[76], the sampling time, the number of windows used and the window-width[93]. Therefore attaining a binding free energy, which compares well to the experimentally determined binding free energy may require further optimization of these parameters for a specific system. Since this is not the scope of this work, we did not perform any further analysis utilizing different parameters set for BN-BS. Further possible explanations for the deviation from experiment could be inaccuracies of the force field, neglect of explicit polarization, conformational changes of the protein that are not captured during ns-scale time simulations etc. We note that previous, shorter simulations by Neumann and Gottschalk[92] as well as by Wang and Helms[91] also resulted in too low PMF profiles.

ΔG_{PMF} contributes the most to the standard free energy of binding because it accounts for all direct interactions, solvent contributions, rotational and translational entropies of the proteins. The term ΔG_v mainly accounts for translational entropy since among the enthalpy and entropy terms translational entropy is the only one that has a concentration dependence[94, 95]. The area explored by the protein in the *xy* plane, 1.61 Å^2, was computed analytically (see methods) and, therefore, is the same for all three systems. Since the unbound length, l_u, is also almost the same for all systems, the free energy change ΔG_v is the same up to second decimal. Because the unbound volume, V_u, is smaller than

the standard state volume, the rescaling free energy term ΔG_v yields a positive contribution (2.5 kcal/mol) to the standard free energy of binding. The free energy to remove the orthogonal constraints, ΔG_R, makes a contribution of about -0.9 kcal/mol for the BN-BS and EIN-HPr systems and of about -1.2 kcal/mol for the CC-CC system. The ΔG_R values computed using different time frames do not differ much (see Table 3.2) presumably due to the exponential averaging.

3.4.4. Determinants of Binding Affinity

The riddle in protein-protein association is that there exist no strong chemical bonds formed between the proteins, yet they do form stable assemblies in aqueous solution. As explanation for these strong driving forces Ben-Naim[96] suggested that solvent-induced interactions make a large, favorable contribution to the standard free energy of binding. Ben-Naim decomposed the binding free energy into the following three terms[96]:

$$\Delta G^o = \Delta G_{T-R} + \Delta U + \delta G \quad (3.17)$$

where ΔG_{T-R} is the contribution to the driving force (ΔG_o) due to the changes in translational and rotational degrees of freedom of all species (the monomers and the complex), ΔU is the energy change for bringing the two proteins from infinite separation to the final configuration of the complex in vacuum, and δG is the solvent induced contribution to the binding free energy.

Figure 3.4(d-f) show the direct interaction energy between the complex partners along the reaction coordinate obtained from full-length windows. As for the entropies, the direct interaction energies were computed from the full-length windows by dividing the reaction coordinate into bins and averaging the energies inside each bin. As expected for protein pairs carrying nonzero net electric charges, the direct interactions are overwhelmingly large compared to the corresponding PMF and always favorable except for EIN-HPr at larger distances. The observed humps in the electrostatic energy along the dissociation path (Figure 3.4d-f) are due to protein rotations and reveal the pronounced dipolar character of the proteins. In the EIN-HPr system the proteins have a negative overall electrostatic charge even though the interfaces are oppositely charged. Therefore, the proteins may adopt configurations with positive overall electrostatic interaction energies even in the early stages of dissociation. Since

these effects are compensated by respective protein-solvent interactions (see below), the strong changes in the electrostatic interactions along the reaction coordinate do not lead to spikes in the corresponding PMF values (Figure 3.4a-c). Even though the PMF and Lennard Jones interactions almost vanish beyond the cutoff, the electrostatic interactions are still very strong at this distance. Thus, there must be some effect, which counteracts these interactions and lowers the PMF to its actual value and converges to zero beyond the cutoff.

By decomposing the standard free energy of binding in the manner described above, we obtained the results listed in Table 3.3. The free energy change ΔG_{T-R}(see Table 3.3) was computed as the sum of translational and rotational entropy terms $-T\Delta S_L^{tr}$and $-T\Delta S_L^{rot}$, which were predicted from the SF approach. The free energy contributions due to the loss of translational entropy upon binding are almost equal; 2.4 kcal/mol for BN-BS and 2.5 kcal/mol for the other two systems. The $-T\Delta S_L^{rot}$ term that accounts for the free energy due to rotational entropy loss upon complexation is almost twice the contribution from the translational entropy loss, namely 4.2 kcal/mol, 5.0 kcal/mol and 4.8 kcal/mol for BN-BS, CC-CC and EIN-HPr, respectively. The rotational entropy contribution to the standard free energy of binding for BN-BS computed here (4.2 kcal/mol) is in quite good agreement with the value reported in a recent study[53] (5.8 kcal/mol), which was calculated by applying a series of orientational restraints. ΔU is simply defined as the difference between minimum and maximum value of the total direct interactions.

Table 3.3. The solvent-induced contribution ($\boldsymbol{\delta G}$) to the standard free energy of binding was derived as the difference of the computed $\boldsymbol{\Delta G_o}$ from full-length windows (40 ns for BN-BS and 10 ns for CC-CC and EIN-HPr) and the direct interactions $\boldsymbol{\Delta U}$ value computed from full-length windows. $\boldsymbol{\Delta S_L^{tr}}$ and $\boldsymbol{\Delta S_L^{rot}}$ are translational and rotational entropy changes upon complexation based on the SF approach. $\boldsymbol{\Delta G_{T-R}}$ is the sum of the two entropic terms $\boldsymbol{-T\Delta S_L^{tr}}$ and $\boldsymbol{-T\Delta S_L^{rot}}$, where **T** stands for the temperature.

Systems	$\Delta G_o(\frac{kcal}{mol})$	$\Delta U(\frac{kcal}{mol})$	$-T\Delta S_L^{tr}(\frac{kcal}{mol})$	$-T\Delta S_L^{rot}(\frac{kcal}{mol})$	$\Delta G_{T-R}(\frac{kcal}{mol})$	$\delta G(\frac{kcal}{mol})$
BN-BS	-12.6	-594.6	2.4	4.2	6.6	575.4
CC-CCP	-9.3	-1178.5	2.5	5.0	7.5	1161.7
EIN-HPr	-8.3	-718.4	2.5	4.8	7.3	702.8

Table 3.3 illustrates that according to the energy decomposition suggested by Ben-Naim[96] the overall solvent-induced contribution to the standard free energy of binding is positive. This contradicts the reasoning of Ben-Naim who expected. ΔU to be in the order of only 0.5 kcal/mol, which is not the case. We note, however, that the systems studied here are charged so that this behavior is quite expected. For the EIN-HPr system, there exist some regions along the reaction coordinate where the total direct interactions are repulsive (see Figure 3.4f) what would lead to a favorable negative δG. However when considering the end points along the association path, the solvent-induced interactions are repulsive for this system as well. Therefore we conclude that the solvent-induced interactions counteract the large and favorable total direct interactions and are repulsive overall. This issue, whether the solvent-induced interactions are attractive or repulsive was addressed before. Based on some model systems, Berne and coworkers[46] found that plates with homogenous hydrophilic and hydrophobic sites give rise to repulsive solvent-induced interactions. However, for plates with large hydrophilic domains, they reported attractive solvent-induced interactions coupled with a dewetting transition. The plates with homogenous hydrophobic and hydrophilic sites resemble hydrophilic protein-protein interfaces quite well, since the hydrophilic protein-protein interfaces do not only bear hydrophilic residues, but also contain hydrophobic residues.

We also computed the rotational entropy along the reaction coordinate using the distribution of the three Euler angles (see methods) and provide the results in Figure 3.5. Our aim here was mainly to assess the convergence of rotational motions. As seen in the figure, the term ΔS^{rot} converges to a constant value at large distances that we defined as unbound region. We also computed the rotational entropy contribution using different angle bin sizes. Although the numbers changed slightly at some regions along the reaction coordinate, the overall change of ΔS^{rot}, is negligible (see Figure 3.6). The slight deviation in ΔS^{rot} at these regions is largely due to the inadequate sampling, since every distance bin contains different numbers of snapshots. If we consider the rotational entropy of binding as the difference between minimum (in the bound region) and maximum (in the unbound region) values we obtain -11.0 cal K^{-1} mol^{-1} for BN-BS -10.7 cal K^{-1} mol^{-1} for CC-CC and -9.7 cal K^{-1} mol^{-1} for EIN-HPr. These entropy changes give $T\Delta S^{rot}$ values of 3.4 kcal/mol, 3.3 kcal/mol and 3.0 kcal/mol free energy changes at 310° K, respectively, which are smaller than the contributions obtained using the SF approach (see Table 3.3).

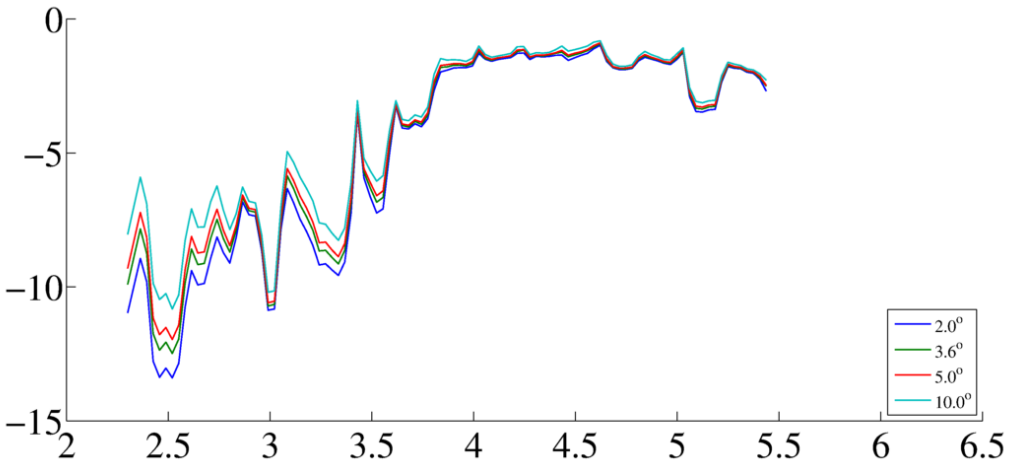

Figure 3.6. Rotational entropy calculated using different angle bin size for the BN-BS complex.

3.4.5. Properties of Interfacial Water

At the atomic level, water appears to be one of the simplest molecules, yet understanding the structure and dynamics of liquid water is an ongoing challenge[97]. Our second set of simulations consisted of MD simulations that address the behavior of water localized between two hydrophilic proteins. It is

well known that confining surfaces can exert a profound influence on the structure and dynamics of water[97, 98]. Here, we investigated how the properties of water localized between two hydrophilic proteins differ from those of bulk water by means of density and orientational order parameter. Figure 3.6 shows that the interfacial water density increases by a few percent at close interfacial distances for the BN-BS and CC-CC systems. However, for the EIN-HPr complex we observed slight dewetting at the same interfacial distances. In all cases, the density reaches the bulk value (0.98 g/cm^3 for TIP3P water model) beyond separation distances of a few nanometers (see Figure 3.7a). A similar behavior was reported previously for protein-protein[48] and model systems[46].

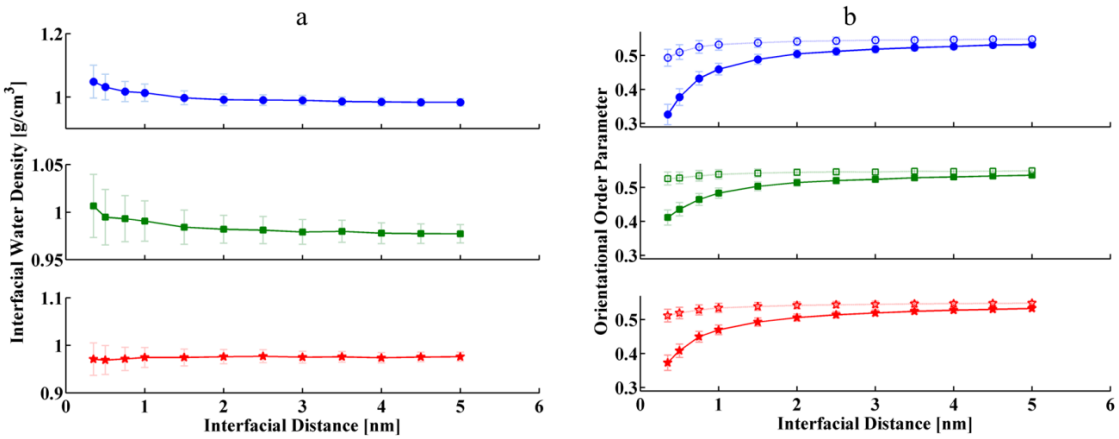

Figure 3.7. Density (panel a) and orientational order parameter (panel b) of the interfacial water for the three systems; Barnase: Barstar (blue circles), Cytochrome *c*: Cytochrome *c* peroxidase (green squares) and N-terminal domain of enzyme I: Histidine-containing Phosphocarrier (red stars). In panel b, the solid symbols are the orientational order parameter when only including neighboring water molecules. The open symbols stand for a modified orientational order parameter of the interfacial water molecules when considering their four closest water oxygens or interfacial protein atoms (O and N) within a 3.5 Å distance cutoff.

In order to investigate the local structure of the interfacial water, we computed the orientational order parameter q. The possible values for this orientational order parameter for a single water molecule range from -3 to 1. However, the average value for a collection of molecules varies between 0 in a random network (no order) and 1 in a tetrahedral network[84]. We conducted this calculation in two different ways; without and with considering the nearby polar

protein atoms (N and O) within a cutoff of 3.5 Å. We found that the average value for the orientational order parameter in bulk TIP3P is 0.55. As shown in Figure 3.7b, the orientational order parameter decreases to values between 0.33 and 0.4 at small separation between the protein surfaces. This was the case for all three systems when we did not consider protein atoms for the analysis. However, when we took into account the nearby polar protein atoms that are potential H-bond donors and acceptors the decrease in the orientational order parameter stops at values of about 0.5. For the CC-CC system the orientational order parameter is even almost constant. We observed the relatively strongest decrease in this modified orientational order parameter for the BN-BS pair (Figure 7b, open circles). The small overall decrease in the orientational order parameter when considering the protein atoms compared to the analysis that only considered neighboring water molecules manifests that most water molecules that are in the vicinity of polar interfaces form hydrogen bonds with the protein atoms as expected.

The small decrease in the orientational order parameter for interfacial water appears to be somehow connected with the increase in density, especially for the BN-BS pair (see Figure 3.8). Structural order decrease with increasing density at constant temperature was previously reported. Errington *et al.*[84] and Yan *et al.*[99] increased the density of water upon compression and observed a decrease in structural order. Another way of looking at this is to consider the electric field created by the confining proteins. The effects of electric fields on water are well documented and are manifold; bending or breaking of hydrogen bonds due to reorientation, phase transition[100], lowering of the dielectric constant[48, 101], etc. have been reported. At small separations the effect of an exerted electric field on the interfacial water molecules is stronger compared to large separations. This is the case because there are only few water molecules in the interfacial gap that can generate a counteracting electric field. The water molecules therefore align with the exerted electric field and do not have enough orientational freedom to rotate into a proper orientation for forming tetrahedral structure.

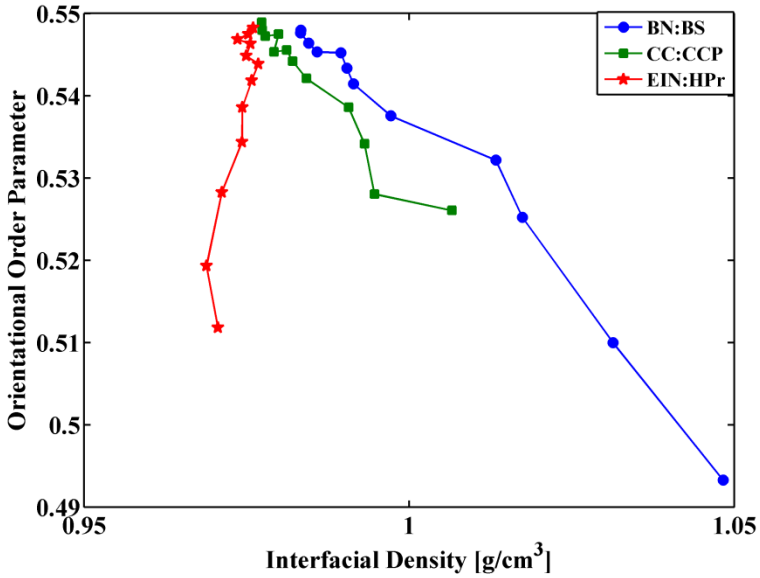

Figure 3.8. Interfacial density vs. orientational order parameter (the polar O and N atoms of the proteins are included) for all three systems.

3.5. Conclusion

In this chapter, we addressed the association of hydrophilic proteins and the role of water using extensive molecular dynamics simulations. This was done on the example of three well studied complexes; Barnase-Barstar (BN-BS), Cytochrome *c* – Cytochrome *c* peroxidase (CC-CC) and the complex of the N-terminal domain of enzyme I with Histidine-containing Phosphocarrier (EIN-HPr). We found that the one-dimensional free energy surfaces of association are downhill or in other words barrierless. However, we note that free energy profiles depend on the reaction coordinate and, therefore, our results do not show conclusively that the real binding paths are barrierless. Using the obtained potential of mean force (PMF) curves along the association path; the standard free energies of binding were computed to be in reasonable to very good agreement with their experimental correspondents. Second, we focused on the role of water in the protein-protein association. Decomposing the standard free energy of binding revealed that the favorable electrostatic and Lennard Jones interactions between the protein pairs render the solvent-induced interactions repulsive. Analysis of the water localized between the two proteins showed that the orientational order parameter of confined water deviates to a small degree but noticeably from bulk values, especially at close separations of the confining

proteins. The water at the interfacial gap is found to be more dense compared to bulk water at close separations of the complex partners. This study showed that different hydrophilic protein-protein interfaces seem to bind according to similar physico-chemical principles. Atomistic MD simulations in explicit solvent proved to be a reliable method to investigate overall principles as well as fine details of such binding processes.

4. How hydrophilic proteins form non-specific complexes

4.1. Summary

In the crowded environment of living cells, proteins frequently encounter other proteins in many possible orientations. Most of these contacts are short-lived because the physico-chemical properties of the two binding patches do not match. However, even for protein pairs that are known to bind tightly, it is by no means an easy task for a protein to find the correct binding site on its partner and align with it to form specific complex. So far the source of interaction specificity that favors a small set of specific "native" interactions over the multitude of alternative orientations is not well understood. As a first step in this direction, we studied in this chapter nonspecific complexes formed by Barnase–Barstar, cytochrome c-cytochrome c peroxidase, and the N-terminal domain of enzyme I- histidine-containing phosphocarrier. Our aim was to characterize structural and energetic aspects of the nonspecific complexes and compare to the native specific complexes. First we employed a set of unbiased MD simulations to obtain two different nonspecific complexes for each system. The chosen nonspecific complexes were analyzed in terms of their interface properties and binding energetics. The one-dimensional free energy profiles of the nonspecific complexes were found to be downhill. Using these one-dimensional free energy profiles, the computed standard free energies of binding are compared to the free energy binding of the correspondent specific complex. Moreover, the characteristic of water confined between two protein partners was studied by means of density and orientational order parameter.

4.2. Introduction

One striking characteristics of living cells is that they are crowded by a large amount of proteins, often in the range of 17% and 35% by weight[102]. The large number of proteins together with their large size dramatically reduces the volume available to proteins to move around the cell interior[103]. Therefore,

many proteins collide and interact at every instance in time. The majority of these interactions exhibit very short lifetime and are insignificant, but some of them lead to formation of biologically functional assemblies through specific recognition. It would be quite detrimental for cells if the functionally important protein encounters would be seriously affected in crowded environments. Thus, there must exist an exquisite fine-tuning of the lifetimes of specific versus nonspecific encounters. So far the source of interaction specificity that favors a small set of interactions over the multitude of possibilities is not well understood.

In the crowded environment of the cell, many possible orientations are available to interacting proteins when they approach each other. It is by no means an easy task for a protein to find the correct binding site on its partner and align with its own binding site to achieve the specific interactions. Kinetic and theoretical considerations suggest that the probability of forming a specific complex can be increased by two mechanisms[104]. The first mechanism is known as reduction in dimensionality. An interaction force keeps the proteins in proximity for a prolonged time as well as in a preferred orientation, allowing a more extensive search of the surface of the partner protein by translational and rotational movements. Nonspecific binding of DNA to proteins is one of the well-known examples[105]. The second mechanism applies only to proteins with a charge dipole. For instance, in the complex of cytochrome c and cytochrome c peroxidase, the search for the binding site is limited by dipolar preorientation of the proteins upon their approach[104]. Thus, this mechanism like the first mechanism leads to a dramatic reduction in the area to be searched prior to any contact formation between the partner proteins.

In chapter 3, taking three well-studied protein–protein complexes as model systems, we studied the association of hydrophilic protein–protein pairs and the role of water[106] by atomistic molecular dynamics simulations in explicit solvent and by umbrella potential simulations. The studied complexes were Barnase–Barstar (BN-BS), cytochrome c-cytochrome c peroxidase (CC-CCP), and the N-terminal domain of enzyme I- histidine-containing phosphocarrier

(EIN–HPr). In this chapter, using similar analysis methods we studied nonspecific complexes formed by the same proteins. Here, our aim is to obtain information about both structural and energetic aspects of the nonspecific complexes and reveal how they differ from their specific counterparts.

The association of the BN–BS pair was addressed extensively in previous studies using, for example, Brownian dynamics (BD) and molecular dynamics (MD) simulations[50-53, 107]. Site-directed mutagenesis[108] analysis and BD simulations[52, 109] of BN-BS system suggest that perturbations in charge distribution outside the native binding surface can modulate the association rate. This reveals the importance of nonspecific interactions in specific complex formation.

Utilizing BD simulations, Northrup et al.[55] studied the diffusional association of the CC–CCP pair. Their findings revealed that favorable electrostatic interactions facilitate long-lived nonspecific encounters that, afterwards, convert to the reactive specific complex. Later on, using paramagnetic NMR spectroscopy, Volkov et al.[110] delineated the conformational space explored by the proteins CC and CCP. They found that the dominant orientation of the protein complex in solution is the same as that observed in the crystal structure. According to their estimation the proteins spend >70% of the lifetime of the complex in the dominant orientation, with the rest of the time spent in dynamic encounter state.

The complex between EIN and HPr has become popular among the researchers who study nonspecific binding. Tang et al.[58] have characterized the nonspecific encounters between EIN and HPr. They demonstrated that even a brief, imperfect collision can mediate the formation of specific complex. Using the paramagnetic relaxation enhancement (PRE) technique, the authors monitored transient nonspecific encounters and mapped their distribution to particular protein surfaces. They proposed that once a nonspecific encounter occurs, HPr can then explore the surface of EIN to find a specific binding pocket. In this way, even protein surfaces not involved directly in the specific binding interface facilitate the assembly of the functional complex. Using data from PRE

and replica exchange simulations, Hummer and co-workers [58] studied the transient encounter complexes in the EIN-HPr association. They reported that, in addition to the specific complex, distinct nonspecific complexes exist as well that account for ~10% of relative population. They pointed out that besides accelerating the binding kinetics by enhancing the rate of success of random diffusional encounters, nonspecific complexes also play a role in protein function as alternative binding modes.

Here, we present the findings from molecular dynamics simulations of nonspecific complexes formed by the proteins of the three systems mentioned above. First we employed a set of unbiased MD simulations to obtain nonspecific complexes. Then we applied a pre-filter to the formed complexes and selected two nonspecific complexes from each system for further analysis, namely the one with the largest contact interface and the one with the longest lifetime. The one-dimensional free energy profiles of the nonspecific complexes were found to be downhill. Using these one-dimensional free energy profiles, the computed standard free energies of binding are compared to the free energy binding of the correspondent specific complex. Moreover, the characteristic of water confined between two protein partners was studied by means of density and orientational order parameter.

4.3. Materials and Methods

4.3.1. Parameterization of Proteins

In this chapter, we conducted three different sets of molecular dynamics simulations for the three protein-protein complexes in explicit solvent. The set-up of the systems and the simulation parameters are identical to those described in chapter 3. In brief, the coordinates for the specific protein-protein complexes were retrieved from the protein databank[59]: Barnase-Barstar (PDB ID: 1BRS[49]), Cytochrome *c* - cytochrome *c* peroxidase (PDB ID: 2PCC[54]) and the amino terminal domain of enzyme I and the Histidine-containing Phosphocarrier protein (PDB ID: 3EZB[57]).

The titration states of titratable amino acids were assigned at physiological pH using the program PROPKA (http://nbcr-222.ucsd.edu/pdb2pqr_1.8/)[60]. All

water molecules in the crystal structure were retained and the placement of additional water molecules in internal protein cavities was tested using the program DOWSER[61] and keeping only those with DOWSER energy below -12 kcal/mol. The Amber force field FF99SB-ILDN[62] was used to model the interactions. Short-range non-bonded interactions were cut off at 1.2 nm distance. Long range electrostatic interactions were treated by the particle mesh Ewald (PME)[63] method. Dispersion correction was applied to energy and pressure. Periodic boundary conditions were applied in all directions. To model the water molecules the TIP3P[64] potential was employed that is typically used together with the AMBER force field. Heme parameters for Cytochrome *c* and Cytochrome *c* peroxidase were the same as in chapter 3, see also [106]. All simulations were performed by the GROMACS package, version 4.5.4[65].

4.3.2. Molecular Dynamics Simulations

The first set of simulations consisted of plain molecular dynamics simulations starting with a pair of unbound proteins to obtain the nonspecific complexes. The initial configurations for this set of simulations were generated by displacing one of the proteins to an interfacial distance of 1.0 nm and rotating it by 90^o, 180^o and 270^o about the x or y axes (see Table 4.1). The interfacial distance was increased when there were clashes after the rotation. For each system, therefore, 7 independent molecular dynamics simulations were conducted including the unrotated configuration. Each simulation was repeated once after assigning different random initial velocities. Therefore for each system 14 simulations were conducted. We also carried out 100 ns-long MD simulations of specific complexes to characterize the behavior of interfacial residues along time.

Each initial configuration was placed in a cubic box of TIP3P water. To mimic physiological conditions 100 mM NaCl was added, including neutralizing counterions. Following an initial energy minimization of 1000 steps of steepest descent, each system was equilibrated in two steps where the heavy atoms of the proteins were restrained. The first step involved 100 ps of MD in the NVT ensemble, maintaining the temperature at 310 K. Protein and nonprotein atoms were coupled separately to temperature baths using Berendsen's weak coupling

algorithm[70]. Subsequently, 100 ps of NPT equilibration were performed, keeping the pressure at 1 bar also using Berendsen's weak coupling method[70]. During data collection, the Nose-Hoover thermostat[72],[73] was combined with the Parrrinello-Rahman barostat[74] to regulate temperature and pressure, respectively. Data collection was completed by 100 ns of conventional MD simulation in the absence of any restraints.

4.3.3. Nonspecific Complex Definition

The nonspecific complex definition used here is arbitrary and employed for practical purposes. We analyzed the 100 ns-long MD simulation trajectories using visual inspection to separate those we observed any single binding event in. Afterwards, we assessed the observed binding events in these trajectories whether these events resulted in nonspecific complexes based on the criteria that we set. We defined a complex as nonspecific based on three main criteria; 1) it should last at least 20 ns, 2) each protein should contribute to the interface with at least 10 residues, each with equal or larger than 80% occurrence frequency as interface residue in consecutive snapshots, 3) The resulted complex should not be native like. The occurrence frequency cut-off was determined by monitoring the interface residues of the specific complexes along 100 ns-long plain MD simulations. We found that, at least 80% of the interface residues maintain the contacts with the interface residues of the partner protein in the specific complexes throughout the simulations. We defined a residue as interface residue when it has at least a single heavy atom within 0.5 Å distance to the partner protein.

For each system, two nonspecific complexes that bear either the longest lifetime or the largest area of the contact interface compared to the rest were selected for further analysis. For each nonspecific complex, the coordinates from the first snapshot of the 20 ns or longer lifetime were taken as initial coordinates for umbrella sampling simulations and for the simulations of water localized between two proteins.

Table 4.1. List of the initial simulations performed to obtain nonspecific complexes. We refer to the nonspecific complexes (NSC) based on their initial configuration and whether they monitored in run 1 or run 2. For instance, NSC-r1-y3 is observed in the first simulation (run1) of the initial configuration y3.

Simulations	Run	Distance (Å)	Rotation axis	Rotation (°)
r1-x1	1	10	-	-
r1-x2	1	15[a]	x	90
r1-x3	1	10	x	180
r1-x4	1	20[b]	x	270
r1-y2	1	10	y	90
r1-y3	1	10	y	180
r1-y4	1	10	y	270
r2-x1	2	10	-	-
r2-x2	2	15[a]	x	90
r2-x3	2	10	x	180
r2-x4	2	20[b]	x	270
r2-y2	2	10	y	90
r2-y3	2	10	y	180
r2-y4	2	10	y	270

a-b) For the EIN-HPr system the distance is increased to 15/20 Å due to steric clashes.

4.4. Results and Discussion

4.4.1. Interfacial characteristics of Nonspecific Complexes

In this work we studied the association/dissociation of nonspecific complexes formed between Barnase and Barstar, cytochrome c and cytochrome c peroxidase and between the N-terminal domain of enzyme I and histidine-containing phosphocarrier. The initial structures of the nonspecific complexes were obtained using unbiased molecular dynamic simulations as explained in the methods section. Out of 14 runs of 100 ns duration for each protein pair, we observed binding in 11, 7 and 13 MD simulations of the BN-BS, CC-CCP and EIN-HPr systems, respectively. Especially the MD simulations that were indicated by

x1 resulted in assemblies that resemble the native complexes. However, 100 ns was not sufficient to observe complexes that differ by less than a few Å from the crystal structure of the correspondent specific complex in RMSD. Out of the complexes formed during the simulations, we considered 4 complexes of the BN-BS, 2 complexes of the CC-CCP and 5 complexes of the EIN-HPr systems as "different enough" from the specific complex (see methods section) so that these conformations likely represent non-specific complexes. For each system 2 nonspecific complexes were selected based on lifetime and the area of the binding interface for further analysis. Figure 4.1 shows the selected nonspecific complexes (shown in blue-orange and blue-silver) together with the specific complexes (blue-red). The larger component of each pair (BN, CCP and EIN are represented with blue) was utilized to superimpose the complexes formed by this pair. As seen in the figure, the nonspecific complexes differ substantially from their specific correspondents in respect of orientation and the protein surfaces involved in binding.

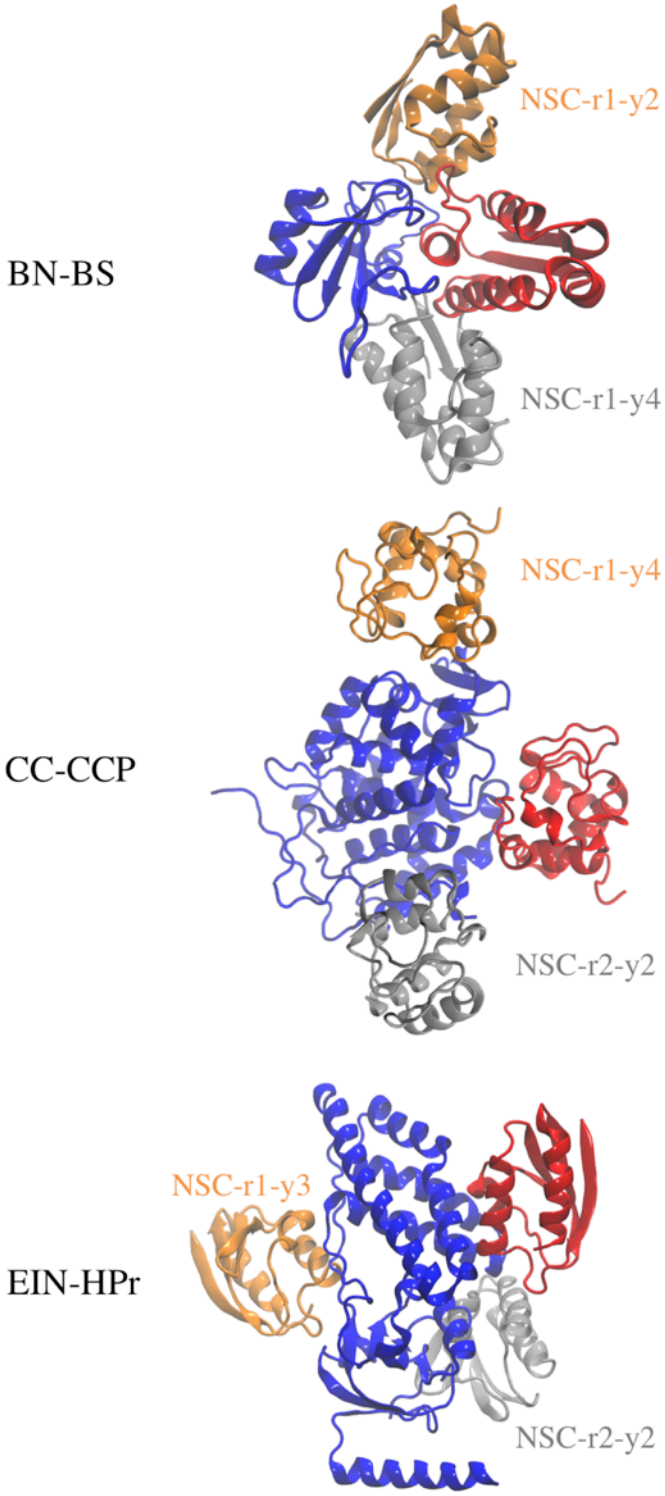

Figure 4.1. Cartoon representation of the nonspecific complexes chosen for further analysis together with their specific correspondents. The larger component of each pair (BN, CCP and EIN are shown with blue) was utilized to superimpose the specific and two nonspecific complexes. The small component (BS, CC and HPr) in specific complex is shown with red and in the nonspecific complexes are shown with orange and silver.

Table 4.2 lists general and interface characteristics of the nonspecific complexes formed. As seen in this table, the areas of contact interface of the nonspecific complexes are smaller than the area of the binding interface of the native complex. This is true for all three systems. For instance, for the BN-BS system the areas of the binding interface of the nonspecific complexes are 478 $Å^2$ and 541 $Å^2$. These values are smaller than the area of the binding interface of the specific complex, which is 763.0 $Å^2$. For the EIN-HPr pair, the contact interface areas of the nonspecific complexes are 470 $Å^2$ and 662 $Å^2$, which is significantly smaller than the area of the binding interface of the specific complex (960 $Å^2$). Whether specific or nonspecific, the complexes formed by the CC-CCP pair bear the smallest areas of binding interface, namely 552 $Å^2$ for the specific, 370 $Å^2$ and 443 $Å^2$ for the nonspecific complexes. The number of interface residues given in Table 4.1 is defined based on occurrence along the trajectory, whereas the area of the binding interface is computed from a single snapshot that was defined as starting structure for the correspondent nonspecific complex. Hence, the area of binding interface comes to the forefront that might be useful for differentiating specific from nonspecific complexes. In relation to this, Hummer and colleagues reported that the area of binding interface of nonspecific complexes constitute 65% of that of the correspondent specific complex[58].

The number of H-bonds and salt-bridges across the binding interfaces differ substantially among the systems. For the BN-BS system, the number of these interactions is substantially lower compared to the specific complex (14 H-bonds and 12 salt-bridges). However in the CC-CCP NSC-r2-y2 complex, the number of H-bonds and salt-bridges exceeds the number of these interactions in the specific complex (4 H-bonds and 2 salt-bridges). All interfaces in nonspecific complexes bear non-zero overall charges except the BN interface in the NSC-r1-r2 complex, because it bears two positively and two negatively charged residues. This indicates the importance of electrostatic interactions, which were reported to be important in nonspecific binding before[58, 104, 111].

Table 4.2. Some global and interface properties of the nonspecific complexes. The number of interface residues calculated from trajectories based on a distance and a frequency of occurrence cut-offs. The interface area, the number of H-bonds and salt-bridges across the interfaces were calculated using the PDBe Pisa database (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) using a single frame.

	Complex	Lifetime [ns]	Area of binding interface (Å^2)	No. of interface residues	No. of H-bonds across interface	No. of salt-bridges across interface	Total charge of interface I : II [e]
BN-BS	Native		763	16 : 14	14	12	+3 : -4
	NSC-r1-x4	20 ns	392	10 : 11	3	0	0 : -2
	NSC-r1-y2	92+ ns	478	15 : 14	1	2	0 : -3
	NSC-r1-y3	31 ns	472	11 : 15	4	3	0 : -5
	NSC-r1-y4	50 ns	541	11 : 11	1	2	+2 : -3
CC-CCP	Native		552	13 : 10	4	2	+5 : -2
	NSC-r1-y4	27 ns	370	10 : 10	4	3	-1 : +1
	NSC-r2-y2	51+ ns	443	11 : 10	6	6	-2 : +4
EIN-HPr	Native		1002	33 : 24	6	5	-5 : +4
	NSC-r1-y3	92+ ns	470	15 : 13	3	5	-3 : +2
	NSC-r1-y4	73+ ns	408	13 : 12	1	0	+1 : +1
	NSC-r2-x3	50 ns	446	13 : 12	8	6	-1 : +3
	NSC-r2-y2	77+ ns	662	17 : 16	3	2	-2 : +3
	NSC-r2-y3	45+ ns	471	10 : 12	8	5	-3 : 0

4.4.2. One Dimensional Free Energy Profile of Nonspecific complexes

By combining umbrella sampling simulations with the Weighted Histogram Analysis Method (WHAM) we obtained the one dimensional free energy profile for the association of the nonspecific complexes as for the specific complexes described in chapter 3. As reaction coordinate we used the center of mass distance between the two proteins. Figure 4.2 depicts the PMF curves of the nonspecific complexes using full-length windows. Even though the values change among the PMF curves, the PMF curves exhibit similar behavior, which is not surprising for the chosen reaction coordinate. Among all nonspecific complexes, the profiles for BN-BS NSC-r1-y2 (Figure 4.1a, blue) and CC-CCP NSC-r1-y4 (Figure 4.1b, blue) converge at 1.0 nm physical separation; therefore they have

the sharpest PMF curves. The PMF curves of the other two nonspecific complexes of these systems converge beyond 1.2 nm physical separation. These distances are shorter than the physical separations that were reported before for correspondent specific complexes[106]. For the EIN-HPr system the picture is somewhat different. For this system, convergence occurs beyond physical separations of 1.4 (NSC-r1-y3) nm and 1.5 nm (NSC-r2-y2), what is comparable to those that we reported before for the specific complexes [106].

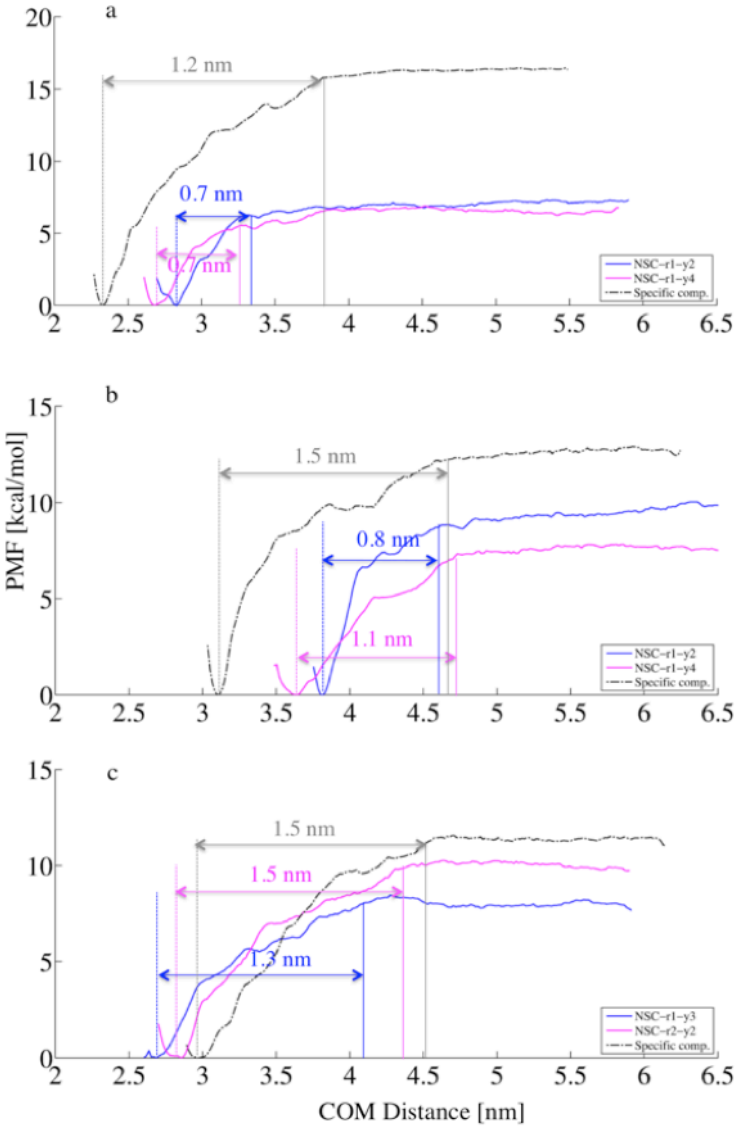

Figure 4.2. Potential of mean force of the nonspecific complexes. a) Barnase: Barstar. b) Cytochrome *c*: Cytochrome *c* peroxidase c) N-terminal domain of enzyme I: Histidine-containing Phosphocarrier. The plain line that is parallel to the y-axis represents the cutoff that separates the bound region from the unbound region. The dashed line left of it marks the position of the bound state.

To obtain insight into the energetic contributions that lead to these PMF profiles, we analyzed the components of the nonbonded interaction terms (see Figure 4.3). For this we divided the reaction coordinate into bins and recorded the average energies inside each bin. The separation distances where the PMFs start to flatten coincide with the distance where the direct Lennard-Jones (LJ) interactions between the proteins almost vanish along the reaction coordinate. However, the electrostatic interactions are still very strong at these separations (see Figure 4.3). Since both proteins carry nonzero net electric charges, the direct interactions are overwhelmingly large compared to the corresponding PMF values along the reaction coordinate. Except for the EIN–HPr system at larger distances, the direct integrations are always favorable. For the EIN-HPr complexes, the direct interactions between the proteins are dominated by charge-charge repulsion at large distances. Although the PMF curves are relatively smooth, the correspondent direct interaction curves have irregularities. The observed bumps and dips in the electrostatic energy along the dissociation path are due to protein rotations and reveal the pronounced dipolar character of the proteins. Due to the compensation of these effects by respective protein–solvent interactions the strong changes in the electrostatic interactions along the reaction coordinate do not lead to bumps and dips in the corresponding PMF values.

For the association of Bn and BS; previous Brownian dynamics simulations[51, 109] as well as unbiased MD simulations[48] provide good evidence that an approach perpendicular to the protein surface is energetically preferred. For the CC:CYP pair, this appears plausible as well given the strong electrostatic complementarity and the clearly preferred native complex[56].

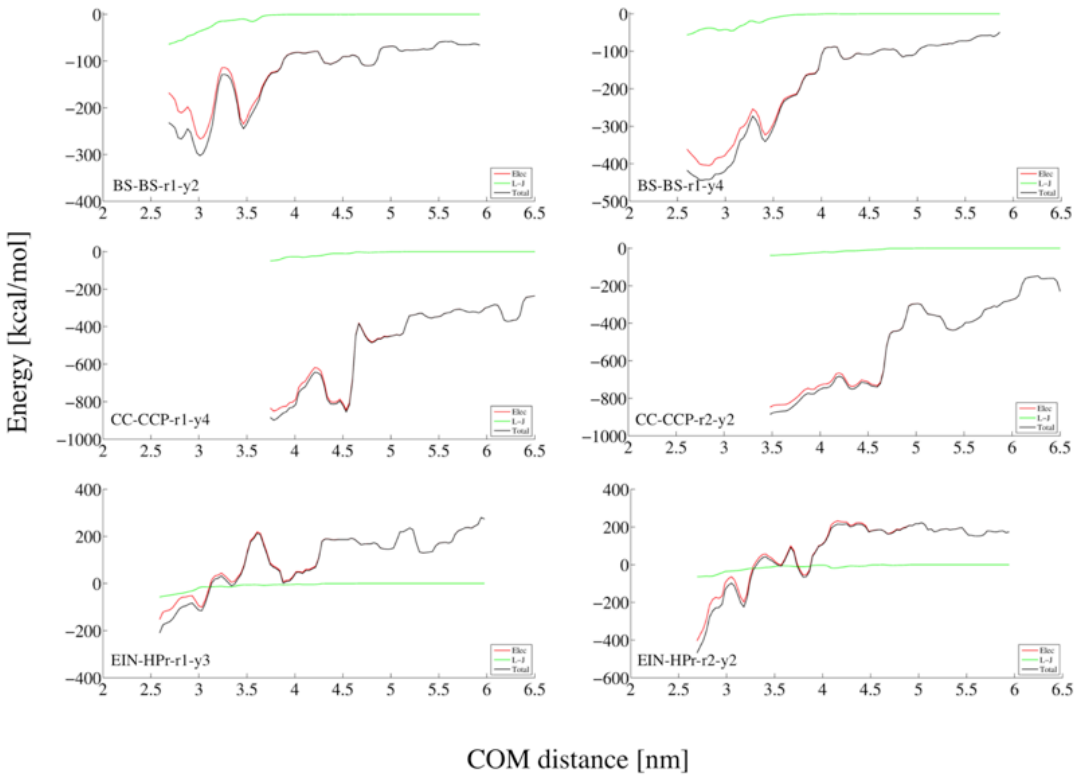

Figure 4.3. Direct interactions between the two proteins calculated from umbrella sampling simulation windows along the COM distance. Barnase: Barstar (upper panel), Cytochrome *c*: Cytochrome *c* peroxidase (middle panel) and N-terminal domain of enzyme I: Histidine-containing Phosphocarrier (lower panel).

4.4.3. Standard Free Energy of Nonspecific Binding

Based on the PMF values given in Figure 4.2, we computed the standard free energies of binding for the nonspecific complexes, as explained in chapter 3. The results are provided in Table 4.3 together with the experimental values for specific binding. The cutoff between bound and unbound regions was set to the value of the reaction coordinate where the PMF becomes constant within some error interval. In Figure 4.2, the cutoff distances are marked as dashed lines parallel to the y-axis. The standard energy of binding for the six nonspecific complexes varies between -3.7 kcal/mol and -7.1 kcal/mol. As expected, the standard free energies of binding computed here for nonspecific complexes are less favorable than the values computed previously for the corresponding specific complexes[106] and the experimental values. The BN-BS system gave the largest difference in binding free energies for specific and nonspecific complexes. The reported experimental value for the BN-BS specific complex is

19.0 kcal/mol, while the computed free energies of binding for the nonspecific complexes studied here are -3.72 kcal/mol and -3.50 kcal/mol. As discussed before, the binding interface area of the specific complex is larger than that of the nonspecific complexes for the BN-BS pair. The interface property that differs most noticeably between specific and nonspecific complexes is the number of H-bonds and salt bridges formed across the interface. The strength of the direct interactions between proteins might be another explanation for the dramatic difference in the binding free energies for specific and nonspecific complexes of this pair. Despite bearing the smallest areas of the contact interfaces, CC-CCP nonspecific complexes do not have the least favorable standard energies of binding. They are -8.8 kcal/mol for the specific, -4.8 kcal/mol and -6.0 kcal/mol for the nonspecific complexes. This can be attributed to the considerable number of H-bonds and salt-bridges across the interface and the relatively strong electrostatic interactions between the partner proteins. The nonspecific complex EIN-HPr NSC-r2-y2 has the most favorable standard free energy of binding among all complexes and this value is close to the standard free energy of binding of the specific complex. The area of the contact interface of this nonspecific complex (662 $Å^2$) is the most prominent interface property. Considering that both proteins are negatively charged, the direct interaction energy between the proteins is quite favorable for this complex, which is comparable to the direct interaction energy calculated for the specific complex[106].

ΔG_{PMF} contributes the most to the standard free energy of binding since it accounts for all direct interactions, solvent contributions, rotational and translational entropies of the proteins. The rescaling term ΔG_v mainly accounts for translational entropy. The terms ΔG_v and ΔG_R do not differ substantially among complexes. Therefore it is ΔG_{PMF} that determines the trend among the standard free energies of binding.

Table 4.3. Standard free energies of binding for the nonspecific complexes. The experimental values $\boldsymbol{\Delta G^{o}_{exp}}$ are based on the binding constants given in [ref]. The calculated values; $\boldsymbol{\Delta G_{PMF}}$ $\Delta \boldsymbol{G_{v}}$ and $\boldsymbol{\Delta G_{R}}$ stand for the free energy change of binding between the bound and unbound regions of the PMF, the free energy change from the unbound volume to the standard state volume and the free energy change to remove the orthogonal restraints, respectively.

		$\Delta G_{PMF}(\frac{kcal}{mol})$	$\Delta G_{v}(\frac{kcal}{mol})$	$\Delta G_{R}(\frac{kcal}{mol})$	$\Delta G^{o}(\frac{kcal}{mol})$	$\Delta G^{o}_{exp}(\frac{kcal}{mol})$
BN-BS	NSC-r1-y2	-5.10	2.42	-1.04	-3.72	-19.6
	NSC-r1-y4	-4.77	2.41	-1.14	-3.50	
	Specific	-14.2	2.5	-0.9	-12.6	
CC-CC	NSC-r1-y4	-7.39	2.35	-0.96	-6.00	-8.8
	NSC-r2-y2	-6.06	2.43	-1.17	-4.80	
	Specific	-10.6	2.5	-1.2	-9.3	
EIN-HPr	NSC-r1-y3	-6.53	2.59	-1.47	-5.41	-7.8
	NSC-r2-y2	-8.57	2.59	-1.12	-7.10	
	Specific	-9.9	2.5	-0.9	-8.3	

As evident from our discussion above, it is difficult to connect the standard free energies of binding and the characteristics of binding interfaces. It is even more difficult to draw a conclusion on what really differs between specific and nonspecific complexes that is common for all pairs. The most obvious result is that nonspecific complexes have less favorable free energies of binding than the native complexes, which is not surprising.

4.4.4. Factors Determining Binding Affinity

Recently, based on a formula suggested by Ben-Naim[45] we assessed the determinants of binding affinity for specific complexes. According to this formula the binding affinity is decomposed into three terms:

$$\Delta G^{o} = \Delta G_{T-R} + \Delta U + \delta G$$

where ΔG_{T-R} is the rotational and translational entropy contribution to the binding affinity (ΔG^{o}), ΔU is the energy change for bringing the protein partners from infinite separation to the final configuration in vacuum and δG is the solvent-induced contribution to the free energy of binding. Table 4.4 lists the energetic and entropic factors determining the binding affinity based on the above formula. The term ΔG_{T-R} is calculated as the sum of translational and

rotational entropy terms $-T\Delta S_L^{tr}$ and $-T\Delta S_L^{rot}$, which were predicted using SF approach[78]. The contributions due to the loss of translational and rotational entropies vary between 2.44 - 2.51 kcal/mol and 4.80 - 4.90 kcal/mol, respectively. That gives an overall free energy change (ΔG_{T-R}) of ~7.3 kcal/mol, which compares very well to the entropic contributions calculated for the specific complexes[106]. The SF approach assumes that in bound state the ligand does not have orientational nor cratic entropies. However, even in very tight complexes, the ligand/protein bears some degree of freedom with respect to its partner. Therefore the SF method overestimates the entropic contribution, ΔG_{T-R}. The reason why the entropy loss is almost the same for specific and nonspecific complexes formed by a protein pair might stem from the neglect of the orientational entropy in the bound state, considering that the unbound state is the same for all complexes formed by this protein pair. However, we find it remarkable that the entropy loss is very close for the complexes formed by different systems.

Table 4.4. The solvent-induced contribution (δG) to the standard free energy of binding was derived as the difference of the computed ΔG_o and the direct interactions ΔU value. ΔS_L^{tr} and ΔS_L^{rot} are translational and rotational entropy changes upon complexation based on the SF approach. ΔG_{T-R} is the sum of the two entropic terms $-T\Delta S_L^{tr}$ and $-T\Delta S_L^{rot}$, where T stands for the temperature.

	$\Delta G^o\left(\frac{kcal}{mol}\right)$	$\Delta U(\frac{kcal}{mol})$	$-T\Delta S_L^{tr}(\frac{kcal}{mol})$	$-T\Delta S_L^{rot}(\frac{kcal}{mol})$	$\Delta G_{T-R}(\frac{kcal}{mol})$	$\boldsymbol{\delta G}(\frac{kcal}{mol})$
BN-BS -r1-y2	-3.72	-244.37	2.44	4.85	7.29	233.36
BN-BS -r1-y4	-3.50	-394.47	2.46	4.83	7.29	383.68
CC-CC -r1-y4	-6.00	-663.26	2.48	4.90	7.38	649.88
CC-CC -r2-y2	-4.80	-737.46	2.45	4.89	7.34	725.32
EIN-HPr -r1-y3	-5.41	-489.40	2.53	4.82	7.35	476.74
EIN-HPr -r2-y2	-7.10	-690.62	2.51	4.80	7.31	676.21

ΔU is simply calculated as the difference between the minimum value in the bound region and the maximum value in the unbound region. As seen in Table 4.4, the direct interactions energies are favorable and very strong for all six nonspecific complexes. However, these values are smaller than the values that were computed for their specific correspondents[106]. The solvent-induced

interactions counteract the large and favorable total direct interactions and are repulsive overall. We addressed this issue whether the solvent-induced interactions are attractive or repulsive for the specific complexes before[106]. The sign and strength of the solvent-induced interactions are mainly determined by the direct interactions between proteins. Therefore the solvent-induced interactions for nonspecific complexes are smaller than the values computed for their specific correspondents[106].

Irrespective of whether the native or a nonspecific contact is formed, the unbound state is unique for the complexes formed by a protein pair. What differs among them is the bound state. Assuming that the systems were sampled sufficiently, the values that are given in Table 4.4 for a protein pair are determined by the bound state. For instance which part of the protein surfaces are exposed to solvent and which parts are deprived from building interactions to solvent plays an important role in the strength of the solvent-induced interactions. Analogously, the direct interactions are dependent on the orientation of binding partners with respect to each other in the complex, as well as the characteristics of the surfaces of the proteins involved in binding.

Above we address the determinants of binding affinity based on the formula proposed by Ben-Naim. Our findings suggest that it is the direct interactions between binding partners that favors the formation of nonspecific complexes studied here. The other factors oppose the overwhelmingly strong direct interactions keeping in mind that we did not assess the role of vibrational entropy in binding. Since the entropic contributions are almost the same, it is the balance between direct interactions and the solvent-induced interactions that determines the binding affinity, which is the main difference between nonspecific complexes and their specific correspondents.

4.4.5. Characteristics of Interfacial Water

In our third set of MD simulations we aimed at assessing the characteristics of interfacial water confined between two proteins. It was shown previously that confining surfaces can exert profound influence on the structure and dynamics of water[97, 98]. Here, we quantified the influence of protein surfaces involved in

binding on the properties of interfacial water by means of density and tetrahedral order parameter. Figure 4.4 shows the interfacial water density along the interfacial distance for all nonspecific complexes. Beyond separation distances of a few nanometers the interfacial water density is bulk like. However, at close separations the water density changes slightly except for the BS-BS-r1-y2 and CC-CCP-r1-y2 nonspecific complexes. For these two nonspecific complexes, even at close separations the interfacial water density is bulk like (0.98 g/cm^3 for the TIP3P water model). We only observed slight dewetting for EIN-HPr-r2-y2 complex at very close distances, for which the interfacial water density drops to 0.95 g/cm^3. For the remaining three nonspecific complexes the interfacial water density increases slightly above 1.0 g/cm^3. We should emphasize that, since we did not observe a pronounced expulsion of water from the interfaces, the hydrophobic effect has no role in the formation of these nonspecific complexes. This is not surprising considering that all binding interfaces studied here are hydrophilic.

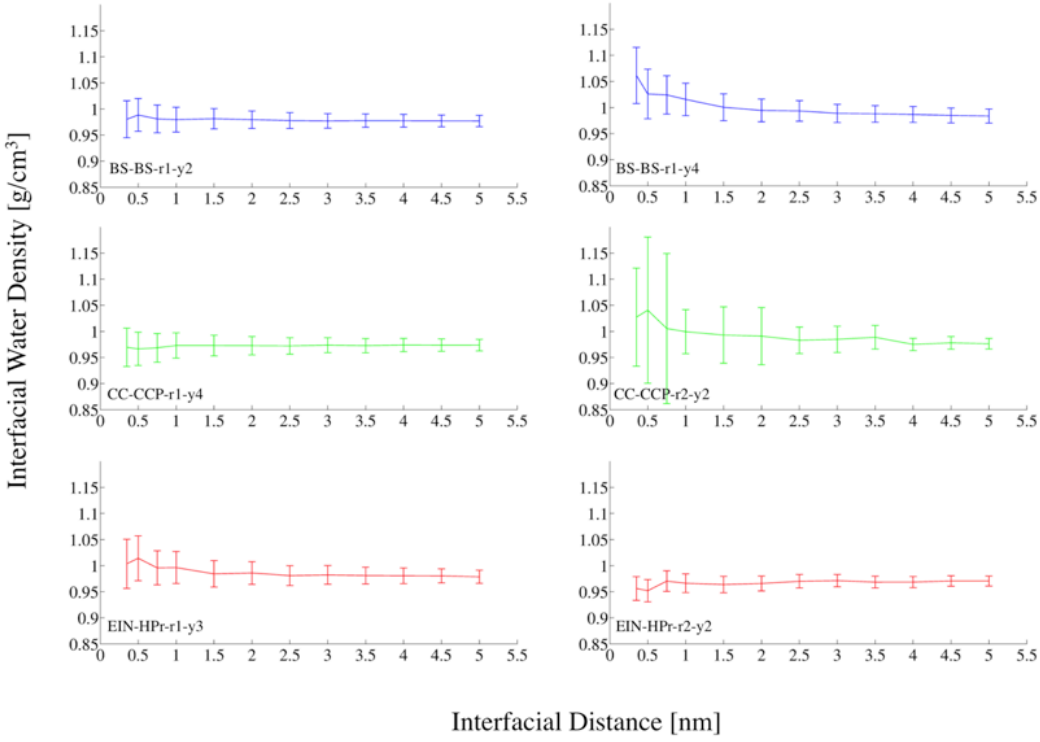

Figure 4.4. Density of the interfacial water for the nonspecific complexes. Barnase: Barstar (upper panel, blue), Cytochrome *c*: Cytochrome *c* peroxidase (middle panel, green) and N-terminal domain of enzyme I: Histidine-containing Phosphocarrier (lower panel, red).

We also computed the orientational order parameter q, in order to quantify the local structure of the interfacial water. The average value of q for a collection of molecules varies between 0 (random network) to 1(tetrahedral network). When calculating the orientational order parameter, the nearby polar protein atoms (O and N) within 3.5 Å as well as water oxygen atoms were considered as potential neighbors. The average value of the orientational order parameter for bulk TIP3P is ~0.55. As shown in Figure 4.5, the orientational order parameter decreases slightly only at close separation distances for all nonspecific complexes. Like the density, it reaches the bulk value beyond separation distances of a few nanometers. A similar behavior was reported by us for the specific complexes (chapter 3 and [106]). Quantifying the water characteristics by means of density and orientational order parameter revealed no real difference in water quality at nonspecific interfaces studied here compared to the specific interfaces studied before.[106] However this might be expected given the fact that whether specific or nonspecific, the studied interfaces are hydrophilic and bear similar physico-chemical characteristics.

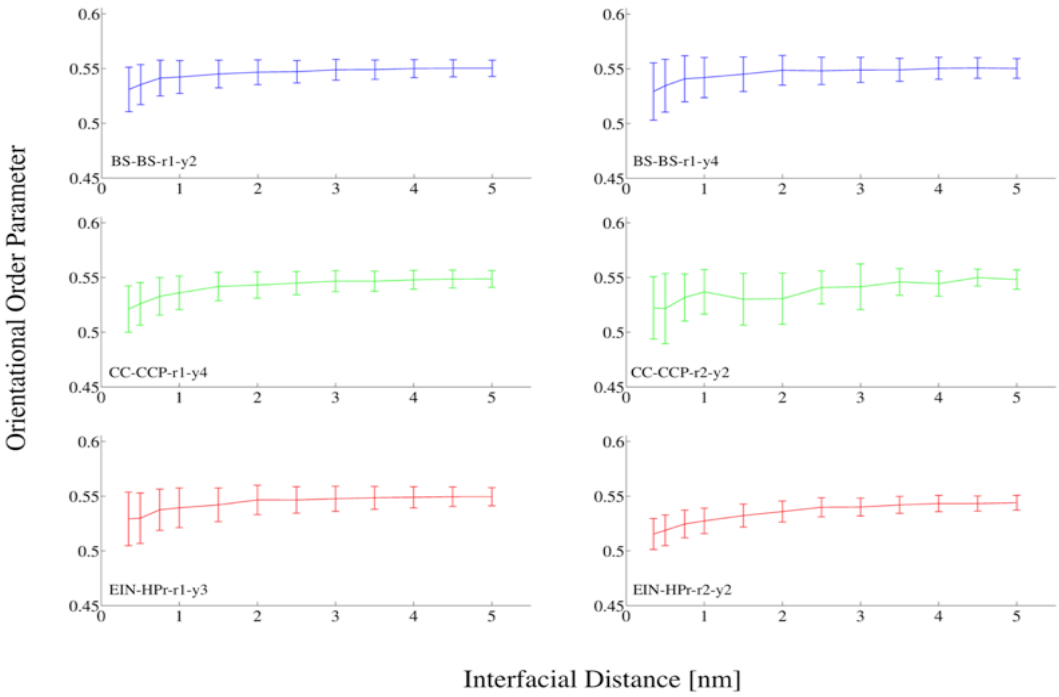

Figure 4.5. Orientational order parameter for the nonspecific complexes. Barnase: Barstar (upper panel, blue), Cytochrome *c*: Cytochrome *c* peroxidase (middle panel, green) and N-terminal domain of enzyme I: Histidine-containing Phosphocarrier (lower panel, red).

4.5. Conclusion

In this chapter, we presented the findings from molecular dynamics simulations of nonspecific complexes formed by three protein pairs. The studied pairs were Barnase-Barstar (BN-BS), Cytochrome *c* – Cytochrome *c* peroxidase (CC-CC) and the complex of the N-terminal domain of enzyme I with Histidine-containing Phosphocarrier (EIN-HPr). Analysis of the interface characteristic of the nonspecific complexes to the correspondent specific interfaces showed that the area of binding interface of the nonspecific complexes is noticeably smaller than the area of binding interface of the native complexes. The one-dimensional free energy profiles of association were found to be downhill or, in other words, barrierless. Using the obtained potential of mean force (PMF) curves along the association path, the standard free energies of binding were computed and compared to the specific correspondents.

Decomposing the standard free energy of binding revealed that it is mainly the direct interactions, both the electrostatic and Lennard Jones interactions that favor the nonspecific protein-protein association. The other factors, namely the entropy of binding and the solvent-induced interactions, oppose the very strong direct interactions. This behavior is the same as what was reported for the specific complexes before. This is not surprising, given the fact that the nonspecific complexes studied here have hydrophilic interfaces and have similar physico-chemical interface properties with the specific complexes.

Analysis of the interfacial water showed that the orientational order parameter of confined water deviate to a small degree but noticeably from bulk values, especially at close separations of the confining proteins. The water at the interfacial gap is found to be more dense compared to bulk water at close separations for some complexes. Even though we observed slight dewetting for one of the nonspecific complexes, the values were within the standard deviation. Therefore we did not monitor a prominent expulsion of water from the interfaces, which is expected considering the residue composition of the interfaces.

We found that specific and nonspecific binding are governed by similar physico-

chemical principles. It might be interesting to study nonspecific complexes that have different interface characteristics than their native correspondents.

5. Quantifying Allosteric Effects

5.1. Summary

Phosphoinostide-dependent kinase-1 (PDK1) is an important protein kinase in insulin and growth factor signaling. PDK1 has an allosteric site termed PIF-pocket, which is located at the N-terminal lobe of the kinase domain. In order to become fully activated, PDK1 docks the phosphorylated conserved hydrophobic motif of substrate kinases at PIP-pocket. Several allosteric allosteric activators of PDK1 have been developed in past years. In order to quantify the allosteric effect, we computed the allosteric coupling energy between the ATP binding pocket and the PIF-pocket of PDK1 as the difference of the binding free energies that were computed in the presence and absence of the allosteric modulator PS182. For this, we designed a thermodynamic cycle. Then, we carried out alchemical free energy perturbation simulations to compute the standard binding free energies utilizing different methods. We found that, the main contribution to the allosteric coupling energy comes from the electrostatic interactions, for the studied system. The contribution due to Lennard-Jones interactions did not change notably based on the method used. We also assessed the role of the initial structures used for the free energy perturbation calculations.

5.2. Introduction

Allostery is a universal phenomenon whereby a perturbation by an effector at one site of the molecule causes a functional change at another site of the molecule via alteration of its shape and/or dynamics. Complex macromolecules such as proteins exist in a multitude of closely related interconvertable conformational states. At any instant, a set of proteins will adopt different conformations that bear similar free energy. These collections are termed ensembles. Allostery is a property of these conformational ensembles, as a perturbation at any site in the structure leads to a shift in the distribution of the conformational states across the entire population. In other words, allosteric perturbations regardless of whether they are structural and/or dynamical do not

create new conformational states; they only change the relative distributions of the states within the ensemble.

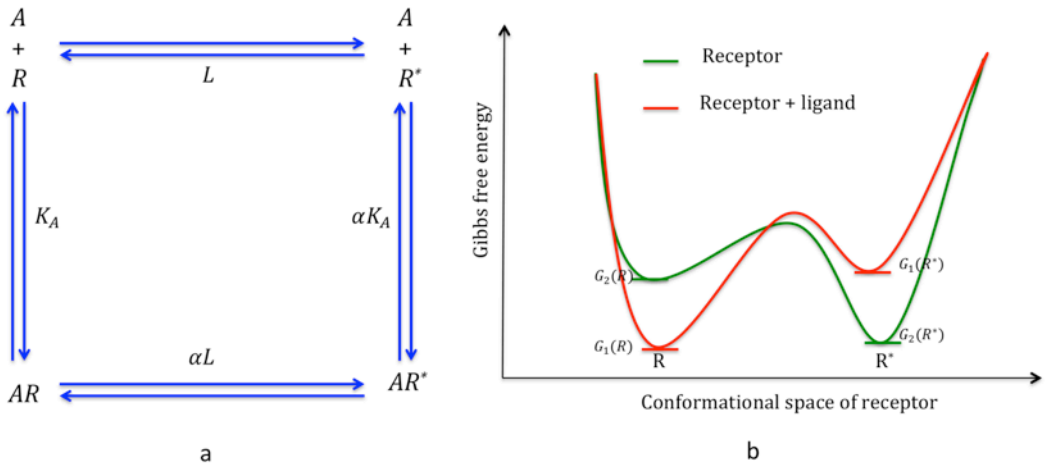

Figure 5.1. a) Two-state model defines allostery as an equilibrium between two states, namely R and R^*, with the equilibrium constant L and their binding to an allosteric ligand A. The equilibrium constants for the inactive and active states are $K_A = [AR]]/[A][R]$ and $\alpha K_A = [AR^*]/[A][R^*]$, respectively. Since all states form a closed thermodynamic circle, the equilibrium constant between AR and AR^* is $\alpha L = [AR]/[AR^*]$. b) Thermodynamic view of allostery. At the bottom of the funnel, the receptor has two populated states: R (active) and R^* (inactive). In the free form (shown in red), the dominated state is the inactive state. Through binding of the allosteric ligand, a population shift occurs in favor of the active state, which dominates in the complex form (shown in green).

Evolution has optimized proteins to populate several switchable states. Each state corresponds to a local free energy minimum on the free energy surface. If a triggering event that effectively switches the protein population from one state to the other is located far away from the active site of the protein, this process is termed allostery. The simplest but quite practical model for the thermodynamic view of allostery is the two-state model. As depicted in Figure 5.5, a protein can populate one of two states: the inactive (R) state and the active state (R^*). In the free form, the distribution of these two states is governed by the equilibrium constant $L = [R^*]/[R]$. The ligand (A) will prefer one state over the other state (conformational selection). The association constants for inactive and active conformations are: $K_A = [AR]/[A][R]$ and $\alpha K_A = [AR^*]/[A][R^*]$, respectively. The binding affinity ratio of ligand A for R* and R defines the allosteric efficacy of

ligand A: $\alpha = \alpha K_A/K_A = [AR^*][R]/[AR]R^*$. The allosteric two-state model offers allosteric ligands a functional classification through the measured allosteric intrinsic efficacy, α. A full agonist can reach nearly 100% activity with $\alpha \gg 1$. A partial agonist corresponds to $\alpha > 1$. A natural antagonist is defined by $\alpha = 1$, where the ligand shows no binding preference, in contrast to an inverse agonist $\alpha < 1$, where the ligand preferentially binds to the inactive state conformation[112].

Before binding of the allosteric ligand occurs, the relative free energy between the inactive (R) and active (R^*) states is given by $\Delta G_1 = G_1(R^*) - G_1(R)$. After binding, the relative free energy between active and inactive states becomes $\Delta G_2 = G_2(R^*) - G_2(R)$. The free energy change due to ligand binding, $\Delta\Delta G = \Delta G_2 - \Delta G_1$, is equal to $RTln\alpha$ [112]. Therefore, the allosteric effect is merely determined by the allosteric efficacy, α, but not by the absolute binding affinity of the allosteric ligand. The allosteric efficacy α is also termed cooperativity factor because it links two structural sites. Here, we should emphasize that allosteric energy transfers is reciprocal, which can be easily derived from energy balance or detailed balance conditions[1]. Consequently, the cooperativity factor α quantifies the effect of an allosteric modulator A on the affinity of a protein to another ligand B. In the same way, it also quantifies the reciprocal effect ligand B has on the affinity of the protein to allosteric modulator A.

Allosteric effects can be quantified experimentally. For instance, one can compare the dose-response curves obtained in the presence of varying concentrations of allosteric modulators to the behavior predicted by quantitative models[113]. Ricci et al.[114] designed allosteric activators and inhibitors for artificial biosensors based on a population shift model. Traditional approaches measure the affinity using a radioligand, i.e. a radioactively marked ligand. Kostenis and Mohr[115] evaluated the modulation of a receptor, using two-point kinetic experiments that determine the delay of the dissociation of the radioligand triggered by allosteric modulation. Homogeneous biochemical fluorescent assays have also been used to characterize ligand affinity and dissociation rates in the presence of allosteric modulators[116].

In this chapter, we used computational tools to quantify the allosteric effect taking place in the enzyme phosphoinostide-dependent kinase-1 (PDK1). PDK1 is a "master" protein kinase in insulin and growth factor signaling. Phospholipid secondary messengers trigger the activation of a complex protein kinase network in which PDK1 is responsible for the activation of many different protein kinases[117]. In order to be activated, the cytosolic PDK1 substrates first need to be phosphorylated at a conserved hydrophobic motif at the C-terminus. This specific sequence binds to the allosteric site on PDK1, termed PIF-pocket (for *PDK1* interacting fragment), that is located at the N-terminal lobe of the kinase domain [118]. The specific binding event enables PDK1 to become fully active and phosphorylate the substrate [119]. Engel *et al.* [120] and Stroba *et al.* [121] showed that the PIF-pocket is a druggable site and developed allosteric modulators of PDK1. Busschots et al. developed two potent compounds (PS182 and PS210) that act as activators of PDK1 and bind to the allosteric PIF-pocket[122] (see Figure 5.4). Using free energy perturbation calculations, we computed the affinity of ATP$(Mg^{2+})_2$ to PDK1 in the presence and absence of the allosteric modulator PS182 (see Figure 5.2). ATP$(Mg^{2+})_2$ binding shifts the pre-existing equilibrium between a binding-competent state and a nonbinding state. In the absence of the allosteric modulator, the binding affinity of ATP(Mg2+)2 to PDK1 is the intrinsic affinity of ATP(Mg2+)2 to the binding-competent state (reaction I in figure 5.2). The binding of the allosteric activator stabilizes the binding-competent state, thus, increases the ATP(Mg2+)2 affinity to PDK1 (reaction II in Figure 5.2). Accordingly, the allosteric coupling energy between the active site (ATP binding pocket) and the allosteric site (PIF-pocket) of PDK1 can be calculated as the difference of the two binding free energies.

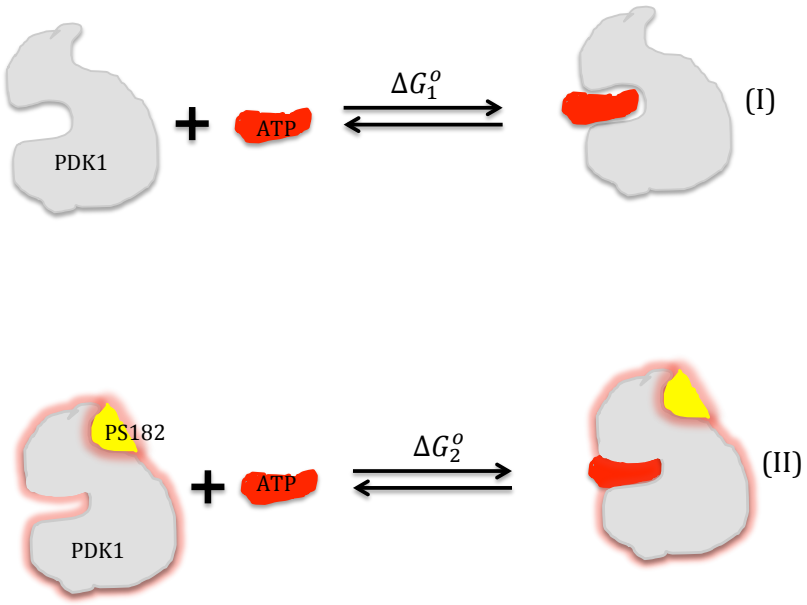

Figure 5.2. In the absence of the allosteric modulator, the binding affinity of ATP(Mg^{2+})$_2$ to PDK1 (ΔG_1^o) is the intrinsic affinity of ATP(Mg^{2+})$_2$ to the binding-competent state (upper panel). The binding of the allosteric activator stabilizes the binding-competent state, thus, increases the ATP(Mg^{2+})$_2$ affinity to PDK1 (lower panel). Accordingly, the allosteric coupling energy between the active site and the allosteric sites is $\Delta\Delta G^o = \Delta G_2^o - \Delta G_1^o$.

5.3. Material and Methods

The coordinates for the protein-ligand complexes were retrieved from the protein databank[59]: PDK1-ATP (PDB ID: 3HRC[123]) and PDK1-ATP-PS182 (PDB ID: 4AW0 [122]).

5.3.1. Parameterization of the Enzyme

The first amino acid (Arg75) and four consecutive amino acids (residues 233-236) in the activation loop were not structurally resolved in the 3HRC coordinate set. In order to complete these missing residues, we superimposed the two coordinate sets with the pymol program and copied the four amino acids from 4AW0 to 3HRC. In the same manner, the missing Mg^{2+} ions were added to the coordinate set 3HRC.

We used the webserver PROPKA (http://nbcr-222.ucsd.edu/pdb2pqr_1.8/)[60] to assign the titration states of titratable amino acids at physiological pH. All crystal water molecules were retained and the placement of additional water

molecules in internal protein cavities was tested using the program DOWSER[61] and only those with DOWSER energy below -12 kcal/mol were kept. For all types of interactions the Amber force field FF99SB-ILDN[62] was used.

In both coordinate sets, the residue Ser241 was mono-phosphorylated. Force field parameter files for phosphoserine in the Amber format were retrieved from the website http://www.pharmacy.manchester.ac.uk/bryce/amber. The file was transferred to Gromacs format using the conversion script amb2gmx[124].

Short range nonbonded interactions were cut off at 1.2 nm distance. For long range electrostatic interactions the particle mesh Ewald (PME)[63] method was utilized. Dispersion correction was applied to both energy and pressure. Periodic boundary conditions were applied in all directions. Water molecules were modeled by the TIP3P[64] water model. All simulations were performed using the GROMACS package, version 5.0.1.[65].

5.3.2. Parameterization of the Ligand

Quantum mechanical calculations for the ligand PS182 were started from its conformation in PDB entry 4AW0. All QM calculations were performed using Gaussian03[67]. For derivation of partial atomic charges, we followed the standardized protocol commonly used in combination with the original AMBER force fields. The geometry was optimized at HF level using the basis set 6-31G* without any restraints. Using the optimized geometry, we obtained the molecular electrostatic potential from the HF/6-31G* electron density of the ligand. Restricted ESP (RESP) charges[68] were computed using the RESP program under Amber Tools in two steps. In the first stage, charge equivalency on chemically equivalent heavy atoms was imposed. In the second stage, the charges of chemically equivalent hydrogen atoms were equated.

Equilibrium values of the bond lengths, angles and dihedrals of the ligand PS182 were taken directly from the optimized structure. The force constants of missing parameters were adopted from the General Amber Force Field (GAFF)[125] by analogy.

5.3.3. Molecular Dynamics Simulations

In this study, we conducted two different sets of molecular dynamics simulations in explicit solvent. First, we performed plain molecular dynamic simulations to obtain equilibrated initial structure for free energy calculations. We conducted 50 ns-long MD simulations for the coordinate set 3hrc, since the missing Mg^{2+} ions and the activation loop of the coordinate set 3hrc were modeled. The atomic fluctuations of the two Mg^{2+} ions computed over the last 10 ns of the MD simulation after superimposing the backbone atoms of the protein were 1.0 and 0.6 Å. For the 4aw0 system, the total simulation time with ligand PS182 was 400 ps. For the 4aw0 system, we also carried out a 50 ns-long plain MD simulation of unbound PDK1 after removing the PS182 compound (We will term this coordinate set $4aw0^{ligand-}$). $4aw0^{ligand-}$ was meant to be an alternative choice for 3hrc.

We combined alchemical free energy perturbation and the Bennett's acceptance ratio method to compute the free energy of ATP together with two Mg^2 ions ($ATP(Mg^{2+})_2$).

In order to generate equilibrated initial structures for the simulations, each system was placed in a cubic box of TIP3P water. To mimic physiological conditions 100 mM NaCl was added, including neutralizing counterions. This resulted in 55 Na^+ and 56 Cl^- for the 3hrc and $4aw0^{ligand-}$ systems, and 53 Na^+ and 52 Cl^- for the 4aw0 system. Following an initial energy minimization of 1000 steps of steepest descent, each system was equilibrated in two steps whereby the heavy atoms of the proteins were restrained using a force constant of 1000 kJ mol-1 nm-2. The first step involved 200 ps of MD in the NVT ensemble, maintaining the temperature at 300 K. Solute and solvent atoms were coupled separately to temperature baths using Berendsen's weak coupling algorithm[70] with a coupling time of 0.1 ps. All bonds were constraint using the LINCS algorithm[71]. Subsequently, 200 ps of NPT equilibration were performed, keeping the pressure at 1 bar also using Berendsen's weak coupling method[70] with a time constant of 1 ps. During data collection, the Langevin integrator was used for temperature control with a friction coeffient of 1 ps^{-1}. The Parrrinello-Rahman barostat[74] was used to regulate the pressure. Equilibration was completed by 50 ns of conventional MD simulation in the absence of any

restraints. For integrating Newton's equations of motion, a stochastic leap-frog algorithm with a time step of 2 fs was used.

The final coordinates at the end of these trajectories were used as starting configurations for alchemical free energy perturbation calculations.

5.3.4. Alchemical Free Energy Calculation Simulations

We used the Double Decoupling Method (DDM) to compute absolute binding free energies of a ($ATP(Mg^{2+})_2$) unit. DDM is based on hypothetical intermediate states in which the interactions between receptor and ligand are turned off alchemically[126]. A problem with this method is that to compute an accurate free energy from the entirely decoupled state to the next-most decoupled state, the ligand is supposed to explore the entire simulation box. This is not a problem in the ligand-only state, since without the ligand the solvent box is isotropic. To overcome this problem with the complex, we attached the decoupled ligand by a spring to the protein by gradually turning this spring on as the ligand is decoupled.

We used a spring constant of 4184 $kJ/mol.nm^2$ in the fully decoupled state. The force constant was zero when the Coulombic intermolecular terms are fully on and was gradually increased throughout the Coulombic decoupling simulations. The spring is attached on one end at the center of mass of $ATP(Mg^{+2})_2$. The other end is a dummy point that coincides with the center of mass of the Cα atoms of the pocket lining residues of the ATP binding pocket. The average distance that was computed from the last 10 ns of the 50-ns long plain MD simulation was taken as equilibrium distance.

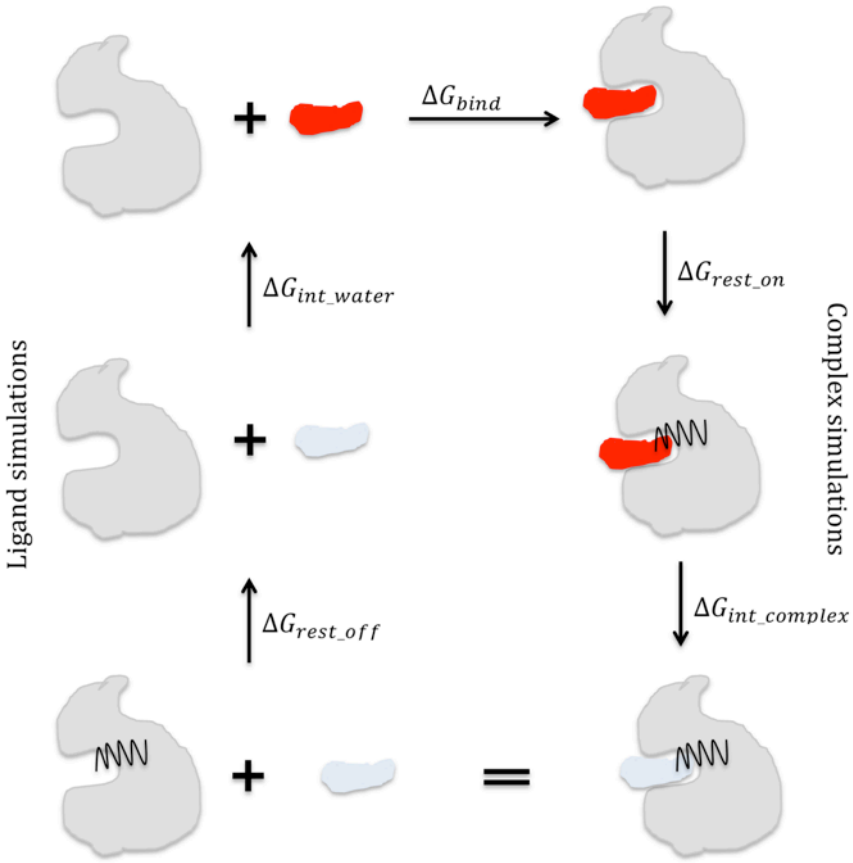

Figure 5.3. Graphical representation of the thermodynamic cycle we followed in this study. The simulations that involve the complex are shown on the right side, whereas the simulations that involve only the ligand are shown on the left side. Before decoupling the ligand from the system, restraints are introduced between the ligand and the protein (top-right corner). Then, the interactions between the ligand and the rest of the system were decoupled gradually (first the Coulombic then the Lennard-Jones interactions). The removal of the restraints is carried out analytically. At this point the ligand comes back to couple with the solvent gradually, which means first the Coulombic then the Lennard-Jones interactions are turned on. In practice, we performed the calculations in the opposite direction. This results in the same free energy with the opposite sign.

Figure 5.3 represents the thermodynamic cycle that we followed, in principle, to calculate the binding free energy of ATP(Mg^{+2})$_2$. The free energy of binding is the sum of the free energy change for the complexation of the ligand and receptor (introduction of the ligand into the binding site) and the free energy of desolvating the ligand (removing the ligand from the solution). Therefore, the free energy of binding is equal to $\Delta G_{bind} = -(\Delta G_{rest\_on} + \Delta G_{int\_complex} + \Delta G_{rest\_off} + \Delta G_{int\_water})$ where $\Delta G_{rest\_on} + \Delta G_{int\_complex} = \Delta G_{complex}$. Then, the

free energy of binding becomes $\Delta G_{bind} = -\Delta G_{complex} - \Delta G_{int\_water} - \Delta G_{rest\_off}$.

We performed the calculation steps shown in the cycle for both the PDK1-ATP(Mg^{+2})$_2$ complex with and without the allosteric modulator PS182 (reactions I and II in Figure 5.2). Since we were interested in computing the difference of the binding free energies ($\Delta\Delta G^o$), we did not carry out the ligand only simulations since the solvation free energy (ΔG_{int_water}) is equal for both systems and cancels out. The restraining free energy (ΔG_{rest_off}) was calculated analytically using the formula $\Delta G_{rest\_off} = -kTln\left(\frac{3}{V}\left(\frac{kT\pi}{K}\right)^{3/2}\right)$ where k is the Boltzmann constant, K is the spring constant, T is the temperature and V is the volume of the reference state (1 particle in 1661 Å^3 = 1 mol/l). Thus, this correction in all cases is 5.3 kcal/mol and accounts for the relative degrees of translational freedom without the need to sample the entire simulation box. Accordingly, the allosteric coupling energy reduces to the difference of the energies that were calculated by turning off the direct interactions (ΔG_{int_comp}).

First we gradually switched off the Coulombic interactions between ATP(Mg^{+2})$_2$ and the rest of the system. For this, the system Hamiltonian was coupled to a coupling parameter $(1-\lambda)$ where $\lambda = 0$ and $\lambda = 1$ correspond to the unperturbed and perturbed states, respectively. Simultaneously we turned the restraining potential on using the coupling parameter λ. The λ values used were: 0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00. No soft-core potential was applied at this step. Afterwards, we turned off the Lennard-Jones interactions in the same manner. During switching off the Lennard-Jones interactions we applied a soft–core potential[127] where soft-core alpha was set to 0.5, the soft-core power to 1.0, and soft-core sigma to 0.3. The same λ values were used. For each λ value an independent simulation was carried out with the same simulation parameters mentioned above. Each simulation was started with 5000 steps of energy minimization using the steepest-descent method, followed by a 200 ps long equilibration in the NVT ensemble and 200-ps in the NPT ensemble with positional restraints applied on heavy atoms. Subsequently, data collection was carried out during a further simulation of 3.5 ns length. Data analysis was

performed with the python scripts pyMBAR (https://simtk.org/home/pymbar) taking the data from last 3.0 ns of the simulations. PyMBAR enables subsampling of the data collected from simulations after calculating the statistical inefficiency, and estimates the free energy by various methods. In this study we used the methods thermodynamic integration (TI), thermodynamic integration with cubic interpolation (TI-cubic), Bennett acceptance ratio (BAR)[28] and multiple-state Bennett acceptance ratio (MBAR)[35]. The uncertainties were computed as described previously[35, 128].

5.4. Results and Discussion

5.4.1. State of the Initial Structures

First, we compared the PDK1 structures that we used for the free energy perturbation calculations (initial structures) with the available PDK1 crystal structures using an approach previously used by Biondi et al.[129] This approach is based on principal component analysis[130] of the crystallographic coordinates. Briefly, this involves the construction of a covariance matrix containing the correlations between atomic shifts with respect to an average structure in the ensemble of all available PDK1 crystal structures with at least 310 amino acids. Diagonalization of this matrix gives eigenvector/eigenvalue sets, which describe concerted shifts of atoms (eigenvectors) together with the corresponding mean square fluctuation of the structures (eigenvalues). This approach allows a condensed description of the conformational states of the initial structures using only a few degrees of freedom. The covariance matrix was built from main chain atoms of residues 77-230 and 245-350. The PDK1 crystal structures were projected onto a subspace spanned by the first two eigenvectors (see Figure 5.5). The active and inactive structures are separated from each other along the first eigenvector. The known most active (close) structures appear as a cluster on the left side of the average structure (has a projection of 0,0) such as 4aw0 (PIF-pocket occupied by PS182), 4aw1 (PIF-pocket occupied by PS210), 4rvv (PIF-pocket occupied by PIFtide) and 4a07 (PIF-pocket occupied by PS171). On the far right side of the graph the inactive structures are located, namely 3qc4 (PDK1 with a DFG-out inhibitor) and 3nax (PDK1 with inhibitor

MP7). As is typical for such two-domain proteins (e.g. lysozyme and protein kinases), the first eigenvector for the PDK1 structures describes a hinge-bending motion between the small and the large lobes in Cartesian space, which causes the opening and closing of the active site. Afterwards, we projected the initial structures onto the eigenvectors obtained from the covariance matrix of the PDK1 crystal structure. As seen in Figure 5.5, the initial structure 4aw0 is the most active-like structure among all structures. This is expected, because the strong allosteric activator PS182 occupies the PIF-pocket of 4aw0. The initial structure $4aw0^{ligand-}$ was meant to be an alternative choice for 3hrc, and was obtained by 50-ns long plain MD simulations after removing PS182. Even though two structures (3hrc and 4aw0) suffice to calculate the allosteric coupling energy in our thermodynamic cycle, we also carried out free energy perturbation calculations for the structure $4aw0^{ligand-}$ to assess the effect of the initial structure.

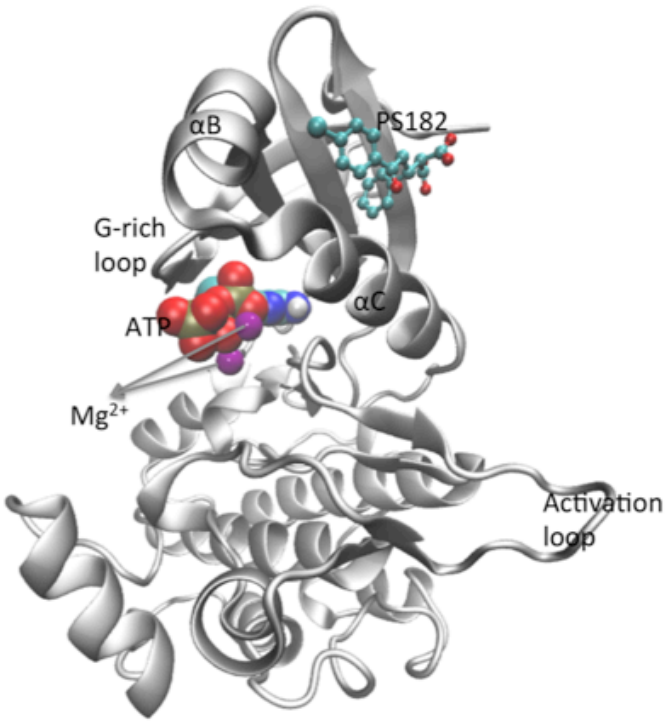

Figure 5.4. Cartoon representation of the PDK1-ATP-PS182 complex (PDB ID: 4AW0). The PIF-pocket and the ATP binding site are indicated by the bound ligands ATP and PS182.

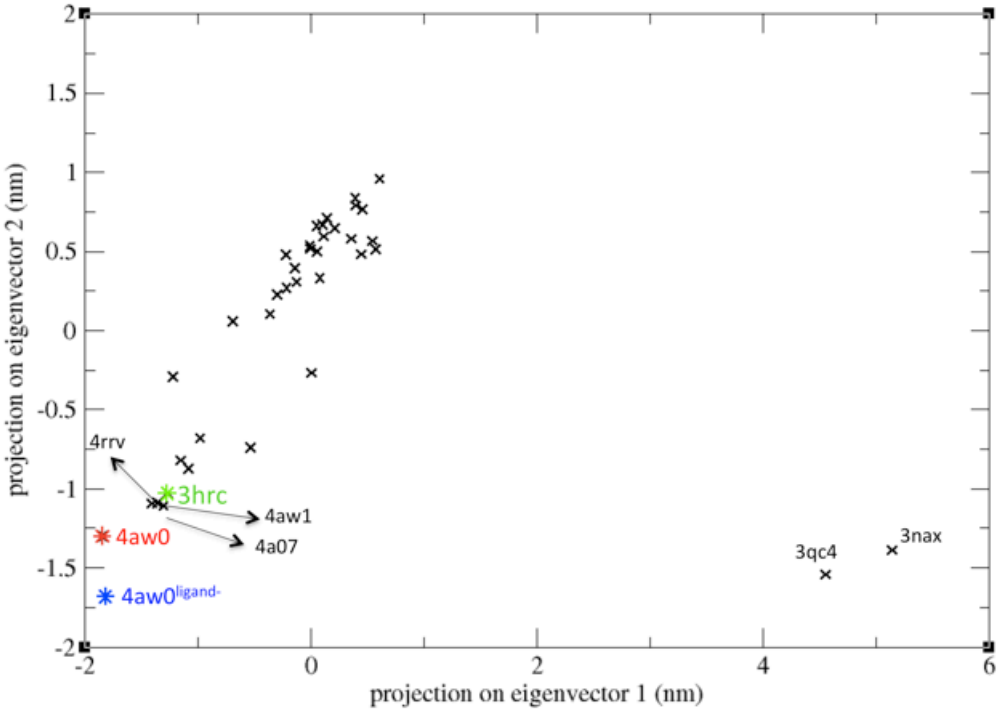

Figure 5.5. Projection of all available PDK1 crystal structures with at least 310 amino acids (cross signs) and the initial structures (color stars) onto the first two eigenvectors obtained from the PDK1 crystal structures. 3hrc (green) represents the structure of PDK1 with ATP(Mg^{2+})$_2$ simulated for 50 ns. 4aw0 (red) is the crystal structure of PDK1 with ATP(Mg^{2+})$_2$ and PS182 equilibrated for 400 ps. 4aw0$^{ligand-}$ (blue) stands for the structure of PDK1 and ATP(Mg^{2+})$_2$ simulated for 50 ns following removal of PS182. For the list of PDK1 structures used for the eigenvector calculation, see Table A.1.

Previous authors have used several specific inter-residue distances as indicator of active vs. inactive conformations of protein kinase A (PKA)[131]. One of those indicators is the distance between residues Ser53-Glys186 that measures the degree of opening of the glycine rich loop. After aligning the respective protein sequences this distance can be measured for PDK1 as well (Ser92-Glys225). It was reported before that binding of an allosteric activator causes the active site Lys111 to approach residue Glu130 in helix αC [123]. Therefore, the Lys111-Gly130 distance is also an indicator for the conformational state of PDK1. Table 5.1 lists these indicator distances for each initial structure. Indeed, the structure with the allosteric activator (4aw0) bears the shortest Lys111 - Glu130 distance (3.5 Å). On the other hand, 3hrc has the most open glycine-rich loop, as expected. The structure 4aw0$^{ligand-}$, which was derived from 4aw0, has a more closed glycine-rich compared to the other two structures. Overall these specific

distances support the picture we obtained from the principal component analysis.

Table 5.1. Indicator distances for PDK1 conformational states. The Lys111-Glu130 distance was measured between the NZ atom of Lys111 and the CD atom of Glu130. The Ser92-Gly225 distance was measured between the $C\alpha$ atoms of the respective residues.

Initial structures	Lys111-Glu130 [Å]	Ser92-Gly225 [Å]
4aw0	3.5	12.2
3hrc	6.3	13.2
$4aw0^{ligand-}$	5.9	11.9

5.4.2. Quantifying Allosteric Effect

We quantified the allosteric effect by computing the allosteric coupling energy as the difference of the binding affinity of $ATP(Mg^{+2})_2$ to PDK1 in the presence and absence of the allosteric activator PS182. For this, we performed alchemical free energy perturbation simulations for three initial structures using a range of the coupling parameter λ. For combining the information from simulations of different λ values to calculate the free energy difference, we used different methods such as thermodynamic integration (TI), thermodynamic integration with cubic interpolation (TI-cubic), Bennett acceptance ratio (BAR) and multiple-state Bennett acceptance ratio (MBAR). Table 5.2 lists the free energy change due to turning on the Coulombic and Lennard-Jones interactions between $ATP(Mg^{2+})_2$ and the rest of the system in the presence of harmonic restraints. $-\Delta G_{int\_comp}$ accounts for the change in Coulombic and Lennard-Jones interactions when the $ATP(Mg^{2+})_2$ is introduced into the binding pocket. As seen easily in Table 5.2 the value for $-\Delta G_{int\_comp}$ correlates well with the activity state of the initial structure. The more active the PDK1 is, the more favorable the free energy of introducing the ligand into its binding site becomes. For instance, 4aw0 is the most active structure (located on the far left of the first principal component) and has the most favorable $-\Delta G_{int\_comp}$ value (-325.7 $\pm$ 0.5 kcal/mol). Similarly, 3hrc is the least active structure and bears the least favorable $-\Delta G_{int\_comp}$ value (-308.3 $\pm$ 0.5 kcal/mol) compared to the other two

structures. The large values of $-\Delta G_{int\_comp}$ are due to the strong Coulombic interactions between the ATP(Mg^{2+})$_2$ and the rest of the system. Even though ATP(Mg^{2+})$_2$ is neutral over all, it bears a strong dipole moment due to the phosphate groups of ATP and the Mg^{2+} ions. We computed the dipole moment of ATP(Mg^{2+})$_2$ as 27.7 D from a 3 ns long MD simulation in solution. For the initial structure 4aw0, doubling the simulations time changes in Coulombic and Lennard-Jones interaction energies by less than 4 kcal/mol and 1 kcal/mol, respectively. Figure 5.6 shows the free energy differences evaluated for each pair of adjacent states for all methods during the discharging and the van der Waals decoupling steps. We calculated the direct interaction energy (Coulombic and Lennard-Jones) between ATP(Mg^{2+})$_2$ and PDK1 from the trajectory of the unperturbed system of 4aw0 using an infinite length cutoff. We found 357.3$\pm$ 21.5 kcal/mol and 24.1 $\pm$ 5.2 kcal/mol for Coulombic and Lennard-Jones interactions, respectively. Additionally, we computed the direct interactions between ATP(Mg^{2+})$_2$ and the allosteric activator PS182 in the same way. The Lennard-Jones interactions turned out to be 0, which is not surprising considering that the distance between the centers of masses of these two ligands was 1.7 nm. At 1.7 nm physical separation, the electrostatic interactions amounted to 2.7 kcal/mol. Therefore, even without the allosteric changes triggered in the binding site, the contribution of the binding of PS182 to PDK1 is 2.7 kcal/mol.

Table 5.3 and 5.4 tabulates the allosteric coupling energies derived from the free energy calculations for the structure pairs 4aw0-3hrc and 4aw0-4aw0$^{ligand-}$, respectively. Expectedly, the allosteric coupling energy is more favorable for the 4aw0-3hrc pair (ranges from -9.7 to -17.5 kcal/mol) compared to the 4aw0-4aw0$^{ligand-}$(varies between -1.5 and -7.4 kcal/mol), since 3hrc is less active (more open) compared to 4aw0$^{ligand-}$. The main contribution to the allosteric coupling energy comes from electrostatic interactions. The contribution due to Lennard-Jones interactions does not change notably based on the method or the structure pair used. However, this is not the case for the contribution due to Coulombic interactions. In comparison to BAR and MBAR, TI and TI-cubic predicts more favorable electrostatic contributions for both structure pairs (see Table 5.3/ 5.4).

Table 5.2. The free energy change due to switching off the interactions of ATP(Mg^{+2})$_2$ in the PDK1 binding site ($-\Delta G_{int\_comp}$) was computed by subtracting ΔG_{rest_on} from the computed free energy of complexation of ATP(Mg^{+2})$_2$ and PDK1 while the restraints were turned on ($\Delta G_{int\_complex} = \Delta G_{complex} - \Delta G_{rest\_on}$). All values are in kcal/mol.

		TI	TI-cubic	BAR	MBAR
4aw0	Coulombic	-310.1 ± 0.5	-309.9 ± 0.5	-307.1 ± 0.3	-308.5 ± 0.3
	LJ	-20.9 ± 0.1	-20.8 ± 0.1	-20.1 ± 0.0	-20.5 ± 0.1
	$-\Delta G_{int\_comp}$	-325.7 ± 0.5	-325.4 ± 0.6	321.9 ± 0.4	-323.7 ± 0.3
3hrc	Coulombic	-294.7 ± 0.4	-294.6 ± 0.4	-299.1 ± 0.3	-299.3 ± 0.3
	LJ	-18.8 ± 0.1	-18.7 ± 0.1	-18.4 ± 0.1	-17.9 ± 0.1
	$-\Delta G_{int\_comp}$	-308.3 ± 0.5	-308.0 ± 0.5	-312.2 ± 0.3	311.9 ± 0.3
4aw0$^{ligand-}$	Coulombic	-305.6 ± 0.5	-305.7 ± 0.5	-307.7 ± 0.1	-307.1 ± 0.1
	LJ	-18.0 ± 0.1	-17.9 ± 0.1	-18.0 ± 0.1	-18.2 ± 0.0
	$-\Delta G_{int\_comp}$	-318.3 ± 0.5	318.3 ± 0.5	-320.4 ± 0.2	-320.0 ± 0.1

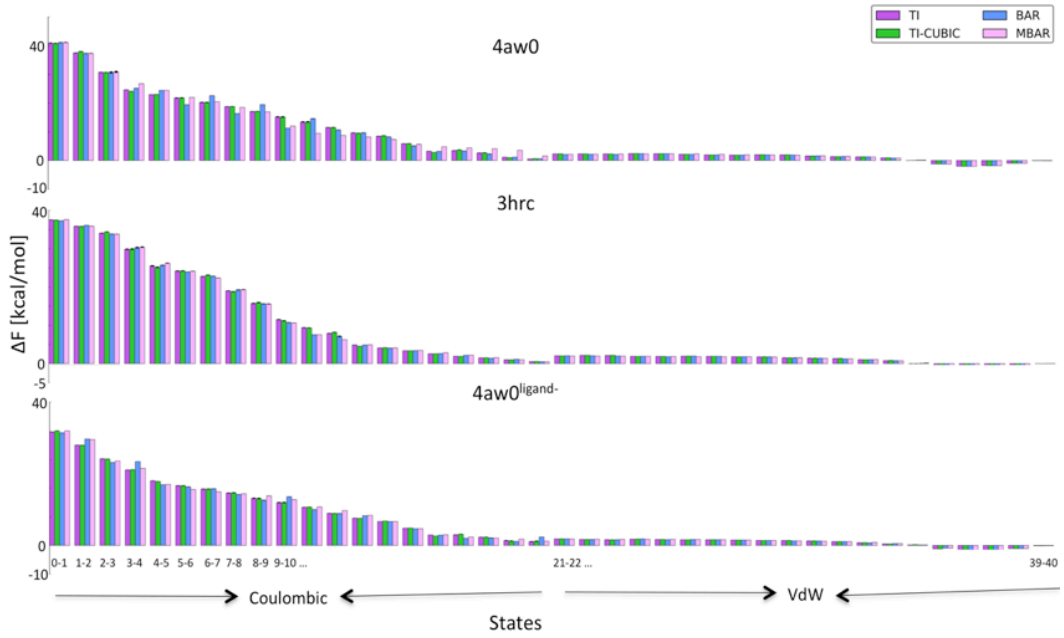

Figure 5.6. The free energy differences evaluated for each pair of adjacent states for all methods, during the discharging (states 0-20) and van der Waals decoupling steps (states 21-40). The system Hamiltonian was coupled to a coupling parameter ($1 - \lambda$). The λ values used were: 0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00.

Table 5.3. Allosteric effect ($\Delta\Delta G^o$) calculated as the free energy difference derived from the initial structures 4aw0 and 3hrc. All values are in kcal/mol.

	TI	TI-cubic	BAR	MBAR
Coulombic	-15.4	-15.3	-8.0	-10.2
LJ	-2.1	-2.1	-1.7	-2.6
$\Delta\Delta G^o$	-17.5	-17.4	-9.7	-12.8

The $\Delta\Delta G^o$ values calculated using different methods differ substantially. Since the uncertainty values listed in Table 1 are very low and close to each other, one cannot assess easily which method performs better. This mainly stems from the large free energy change due to turning on the Coulombic interactions between $ATP(Mg^{2+})_2$ and PDK1.

Table 5.4. Allosteric effect ($\Delta\Delta G^o$) calculated as the free energy difference derived from the initial structures 4aw0 and 4aw0ligand-. All values are in kcal/mol.

	TI	TI-cubic	BAR	MBAR
Coulombic	-4.5	-4.2	0.6	-1.4
LJ	-2.9	-2.9	-2.1	-2.3
$\Delta\Delta G^o$	-7.4	-7.1	-1.5	-3.7

5. 5. Conclusion

In this chapter, we used alchemical free energy perturbation calculations to quantify the allosteric effect in PDK1. We designed a thermodynamic two-state model to capture the phenomenon allostery. According to this model, there exists equilibrium between a binding-competent state and a nonbinding state of PDK1. $ATP(Mg^{2+})_2$ binding to PDK1 shifts this pre-existing equilibrium in favor of the binding-competent state. Similarly, the binding of the allosteric activator stabilizes the binding-competent state, thus, increases the $ATP(Mg2+)_2$ affinity to PDK1. On this basis, we calculated the allosteric coupling energy between the ATP binding pocket and the PIF-pocket of PDK1 as the difference of the two

binding free energies computed in the presence and absence of PS182.

First, we assessed the states of the structure that we used for the free energy calculations using a principal component analysis method and several indicator distances. The structure 4aw0 was found to be the most active structure among all crystal structures. The same analysis showed that 3hrc is an intermediate structure like most of the PDK1 crystal structures. The initial structure $4aw0^{ligand-}$ is quite close to the structure 4aw0 but is less active. Second, we conducted alchemical free energy perturbation calculations for these three systems and analyzed the data using different methods. We found that the allosteric coupling energy is more favorable for the 4aw0-3hrc pair (ranges from -9.7 to -17.5 kcal/mol) than for the 4aw0-$4aw0^{ligand-}$ pair (varies between -1.5 and -7.4 kcal/mol). The main contribution to the allosteric coupling energy comes from electrostatic interactions. The contribution due to Lennard-Jones interactions does not change notably based on the method or the structure pair used. However, this is not the case for the contribution due to Coulombic interactions.

6. Modeling Protein-Peptide Interactions using Molecular Docking

6.1. Molecular Modeling

Molecular modeling encompasses theoretical and computational methods devised to mimic the structural behavior of molecules and molecular processes. Even though it is feasible to use other means to perform molecular modeling studies, computational calculations unequivocally constitute the core of molecular modeling[132]. Molecular modeling techniques are heavily used, for instance, in computational chemistry, computational biology, drug discovery and material science. Below, we will explain the technique molecular docking.

6.1.1. Molecular Docking

Molecular docking is an important technique in structural biology and computer-aided drug design. In molecular docking, one predicts the structure (structures) of the intermolecular complex formed between two or more molecules. Most docking algorithms are able to generate a large number of possible structures; therefore, they require also a means for scoring the structures to identify those of lowest (free) energy that are typically of most interest. Thus, molecular docking is concerned with two main problems, namely the generation and energetic evaluation of the conformations of plausible complexes.

In molecular docking, an important issue is how the receptor and the ligand are represented. The most widely used representations are atomic, surface and grid representations. Due to the complexity of the atomic representation, it is generally used in the ranking phase of docking in conjunction with a potential energy function. Surface-based docking programs are typically used in protein-protein docking studies. The majority of the docking programs use the grid representation of the molecules, especially for the receptor[133].

The simplest version of docking is rigid docking, where one does not take into account the internal dynamics of the molecules to be docked. This form of docking involves six degrees of freedom of one molecule with respect to the other molecule. In flexible docking, additionally, one has to take into account the

internal conformational degrees of freedom of each molecule. Numerous algorithms have been developed to tackle the docking problem. The simplest algorithms treat both molecules as rigid and explore the six translational and rotational degrees of freedom. The well-known algorithm DOCK is a good example for such algorithms[134]. The DOCK algorithm is based on a high degree of shape complementarity between the binding site and a molecule. Most of the methods, which perform flexible docking, consider the conformational space of the ligands and treat the receptor as rigid. Almost all of the known methods for searching the conformational space are incorporated in a docking algorithm. Algorithms that are utilized to treat the molecule flexible can be classified in three groups: systematic methods (incremental construction, conformational search), stochastic methods (Monte Carlo, genetic algorithms, tabu search) and simulation methods (molecular dynamics, energy minimization).

Below we will discuss the widely used docking program AutoDock in terms of the algorithms and scoring functions included. After the detailed discussion of the program, we will give two fruitful applications that we carried out in collaboration with experimental groups.

6.2. AutoDock

Autodock is an automated procedure that is devised for predicting the interactions of ligands to macromolecules. In any docking calculation two conflicting requirements must be balanced: a robust and accurate procedure and a reasonable computational demand. AutoDock is an effort to fulfill both demands. For this, Autodock combines two methods, namely a rapid grid-based energy evaluation and efficient search of torsional degrees of freedom. As we indicated before, molecular docking faces two main problems, which are the generation and evaluation of conformations. Autodock generates the structures using the Lamarckian Genetic Algorithm and utilizes an empirical free energy scoring function for the evaluation of the generated structures. We should stress that the Lamarckian Genetic Algorithm is not the only algorithm available for the conformational search in AutoDock but is the most efficient one.

6.2.1. Conformational Search in AutoDock

As mentioned, the primary method for conformational searching with AutoDock is a Lamarckian genetic algorithm (LGA)[135], which has an enhanced performance relative to simulated annealing and Genetic Algorithm (GA) alone. LGA is a hybrid of a Genetic Algorithm and an adaptive local search method.

The vast majority of the genetic algorithms mimic the hallmarks of Darwinian evolution and Mendelian genetics (one way transfer of information from the genotype to phenotype). In molecular docking, *state variables* of the ligand are given by a set of values that describe the translation, orientation and conformation of the ligand with respect to the receptor. In GA, each state variable corresponds to a gene. The state and the atomic coordinates of the ligand correspond to genotype and phenotype respectively. In the case of molecular docking, the *fitness* is the total interaction energy of the ligand with the receptor. New individuals inherit genes from either parent based on the process called *crossover*, which enables mating of random pairs of individuals. Additionally some offspring undergo random *mutations*. The offspring of the current g*eneration* are selected based on their fitness. Therefore the individuals that are better suited to their environment are kept and poorer suited ones are eliminated.

In earlier versions of AutoDock, optimized variants of simulated annealing were used. Simulated annealing bears both local and global search aspects. Simulated annealing performs a more global search when higher temperatures allow transitions over energy barriers and later on carries out local search when low temperatures enable more focus on local optimization. AutoDock, from version 3 on, has the option of using GA for global search and a local search (LS) method to perform energy minimization, or a combination of both. The LS method is based on the work of Solis and Wets[136], which does not require the gradient information about the local energy landscape when performing the torsional space search. The so-called Lamarckian genetic algorithm is the hybrid of GA and the local search method that we mentioned above. The “Lamarckian” aspect is the feature that enables individual conformations to search in their local conformational space and then pass this information to later generations. In

other words, any environmental adaptation of the ligand acquired during the local search is inherited by its offspring.

6.2.2. Evaluation of the Generated Structures

AutoDock uses a semi-empirical free energy force field to evaluate the structures generated during docking simulations. The force field was parameterized using a large number of protein-inhibitor complexes. The force field evaluates binding in two consecutive steps. In the first step, the change in intramolecular energies upon transition from the unbound state to the conformations of the ligand and protein in the bound state are estimated. The intermolecular energetics upon binding is evaluated in the second step. The force field includes six pairwise evaluations (V) and an estimate of conformational entropy loss (ΔS_{conf})[135, 137]:

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right)$$

where L refers to ligand and P refers to protein in a ligand-protein docking calculation. The first two terms are intramoleculer energies for bound and unbound states of the ligand. Likewise, the following two terms stand for intramolecular energies of bound and unbound states of the protein. The change in the intermolecular energy between the bound and unbound states is given by $(V_{bound}^{P-L} - V_{unbound}^{P-L})$. The two molecules are assumed to be sufficiently distant from each other and, therefore, the intermolecular energy in the unbound state (V_{bound}^{P-L}) is assumed to be zero.

The evaluation for dispersion/repulsion, hydrogen bonding, electrostatics and desolvation consists of the pairwise terms:

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij})^2}$$
$$+ W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{-r_{ij}^2 / 2\sigma^2}$$

The first term is a typical 6/12 potential for dispersion/repulsion interactions.

The A and B parameters were taken from the Amber force field. The second term is a directional hydrogen bonding term and based on 10/12 potential. The directionality of the hydrogen bond interactions $E(t)$ is accounted for by the angle t. Electrostatic interactions are treated by a screened Coulomb potential. The final term is a desolvation potential, which is based on the volume of the atoms surrounding a given atom weighted by a solvation parameter and an exponential term based on distance.

The weighting constants W are optimized to calibrate the empirical free energy. For this task, experimentally characterized complexes were used.

Desolvation is evaluated based on the work by Wesson and Eisenberg[138]. Two pieces of information, namely S_i and V_i, are needed for the evaluation of the desolvation. The S_i values are the atomic solvation parameters that estimate the energy required to transfer the atom from a fully solvated state to a fully buried state. V_i are estimates for the amount of desolvation when the ligand is in the complex.

The term for the loss of conformational entropy upon binding is calculated as follows:

$$\Delta S_{conf} = W_{conf} N_{tors}$$

where N_{tors} stands for the number of torsional angles.

6.3. Application I: The Interaction of BIP with Loop 7 of Sec61

This project was carried out in collaboration with the group of Prof. Richard Zimmermann (Medical Faculty of Saarland University). The modelling results together with extensive experimental results were published in EMBO Journal (2012) 31, 3282-3296.

6.3.1. Background

In mammalian cells, the Sec61 complex mediates the signal peptide dependent protein transport into the endoplasmic reticulum. It is also used to integrate nascent protein sequences translated by the ribosome into the membrane. The gating of Sec61 is tightly regulated due to its role in Ca^{2+} flux.

It is known that the ER luminal binding immunoglobulin protein (BIP) and cytosolic Ca^{2+}-Calmodulin facilitate the gating of Sec61 from the open to the closed state. It was shown by in-vivo studies that BIP together with its nucleotide exchange factor Grp170 plays a role in opening Sec61 for polypeptide passage[139] and together with an undefined co-chaperon functions in closure of Sec61 before and early in translocation[140, 141].

The main goal of this study was to demonstrate the role of BIP in channel closure and reveal the underlying mechanism. Our task in this work was to analyze the putative interactions of BIP with loop 7 of Sec61 using molecular docking.

6.3.2. Material and Methods

We performed flexible peptide docking using the program AutoDock version 4.0[137]. As the optimum length of peptides binding to the human Sec61 protein was determined as 7, we constructed heptapeptides that covered the full length of loop 7 (residues 312-358). All 41 peptide sequences were docked into 14 X-ray structures of bovine Hsc70, DNAK of *G. kaustophilus* and DNAK of *E. coli* and the best model of human BIP that was obtained from multi template homology modeling utilizing the software MODELLER[142]. The PDB IDs of the crystal structures are: 1YUW, 2V7Y, 3KHO, 1DKZ, 1DKY:B, 1DKX, 3DPQ:A, 3DPQ:B, 3DPQ:E, 3DPQ:F, 3DPP:A, 3DPP:B, 3DPO:A, and 3DPO:B. The structural templates used for the homology modeling were 1YUW, 2KHO and 2V7Y. For each structure, ten conformations were generated corresponding to ten docking runs.

6.3.3. Results and Discussion

We found that peptides randomly positioned in a grid centered on the binding channel did never enter into the channel during the docking run. When started from the center of the binding channel, between 0 and all 10 conformations remained bound in the channel. The best docking score obtained was -10.6 kcal/mol. When averaging over the 14 crystal structures plus one homology model, heptapeptides including residues 339-345 (GGLCYYL) and 347-353 (PPESFGS) had average docking scores below -8 kcal/mol.

For comparison, we also performed redocking of the peptide sequences

contained in the respective crystal structures. In each case we only used the seven residues positioned in the channel. This gave optimal scores ranging from -7.3 kcal/mol to -10.8 kcal/mol. Here, the randomly positioned peptides also did not manage to enter into the binding channel of their correspondent binding partners.

This analysis shows that two heptapeptides taken from loop 7 of Sec61 bound to Hsp70 proteins with similar binding affinities as compared to native Hsp70 substrate heptapeptides. This is remarkable given that redocking into co-crystal structures usually results in more favorable docking scores than docking into structures that were crystallized in the presence of other peptides. These results obtained from molecular docking correlated well with the findings from peptide binding experiments. In the peptide-binding experiments performed by our colleagues in Prof. Zimmermann's group, peptides including amino acid residues 329-343 and 339-353 were found to preferentially bind to BIP[143]. This shows the power of molecular docking when combined with experiments.

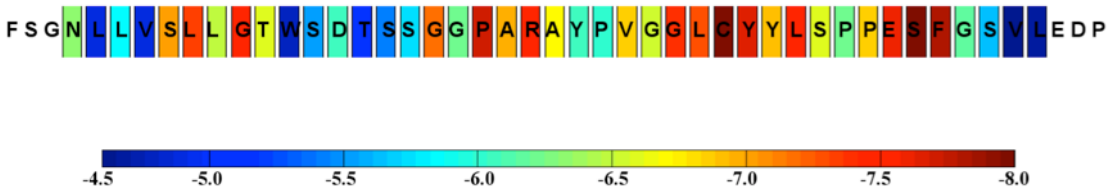

Figure 6.1: Loop 7 is shown in single letter code and with the calculated BiP-binding scores of the respective heptapeptide (kcal/mol). Each heptapeptide is represented by its middle residue.

6.4. Application II: Docking of cyclic peptides that contain non-natural amino acids to CK2

This project was carried out in collaboration with Karsten Niefind (Department of Chemistry, University of Cologne) and his colleagues. The modelling results together with extensive experimental results were published in Analytical Biochemistry (2015) 468, 4-14.

6.4.1. Background

Casein kinase II (CK2) is a highly conserved Ser/Thr kinase composed of two catalytic α-subunits (CK2α) and two regulatory β-subunits (CK2β). The enzyme has been linked to several human pathologies including cancer. Even though CK2 is not an oncogene product, it is dys-regulated by tumor cells due to its anti-apoptotic role. For this reason, CK2 is emerged as a promising pharmacological target and a number of ATP-competitive inhibitors were developed[144]. In order to avoid undesirable binding with off-target kinases several pharmacological strategies have been proposed for CK2. One of the most promising strategies is to perturb the CK2 α /CK2 β interaction by small compounds to alter the substrate specificity of the enzyme. Using a structure-based design methodology and a screening strategy, Laudet and colleagues identified several CK2β-competitive lead compounds [145, 146]. One of these compounds was a cyclic peptide (Pc) of the sequence GCRLYGFKIHGCG that mimics the C-terminus of CK2β namely the interaction region with CK2α. Raaf et al. [147] reported the first X-ray structure of CK2α with this cyclic peptide derived from the C-terminal CK2β segment (see Figure 6.2). They showed that the peptide binds with submicromolar affinity to CK2α, stimulates its catalytic activity and reduces effectively the CK2β binding to CK2α. The results provided a thermodynamic and structural rationalization of the peptide's CKβ - competitive functionality and cleared the way for a peptidomimetic drug perturbing the CK2α/CK2β interaction.

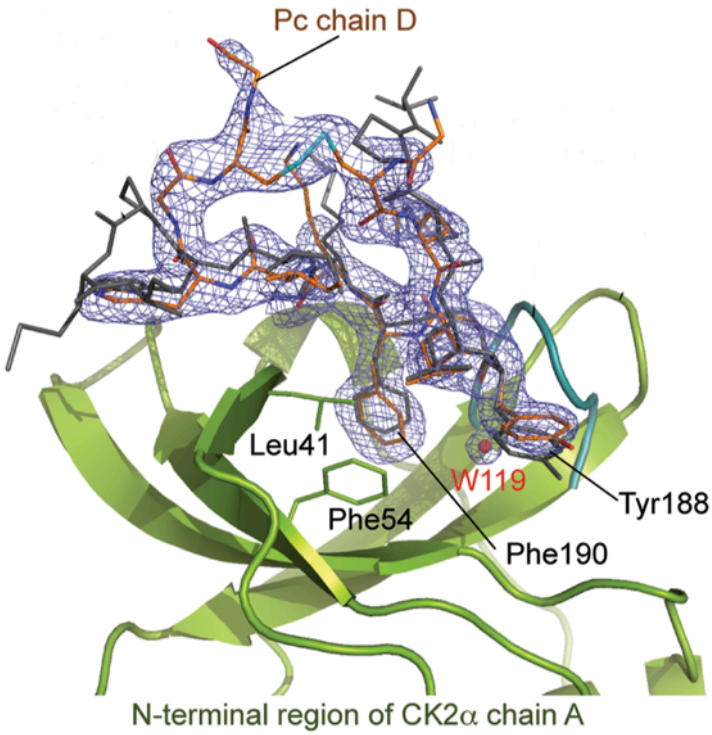

Figure 6.2. Partial view of the CK2α /Pc complex. Highlighted are the Pc chain D and the buried water molecule W119 (W544 here) at the CK2β interface of CK2α chain A (green). The picture was modified from reference [147].

Together with Dr. Karsten Niefind and his colleagues at Chemistry Department in University of Cologne, we used Pc as a scaffold to design new CK2β competitive compounds. In this joint project, we used molecular modeling tools to find potential cyclic peptides that include non-natural amino acids and rank them for further investigations.

6.4.2. Materials and methods

Seven cyclic peptides where Phe-190 was substituted by an non-natural amino acid were docked against the CK2 structure 4IB5 with the program AutoDock version 4.2 [137]. In each docking experiment, the receptor protein, the cyclic backbone of the peptide, and the side chains of the peptide except Phe-190 or its substituents were kept rigid. Heavy atoms of the protein that are not detected in the X-ray structure and protein hydrogen atoms were added using the tleap program of Amber Tools 1.3 [148]. Hydrogen atoms of the non-natural amino

acids in the peptide were added using the program Discovery Studio Visualizer [149]. The input files were prepared using Autodock Tools version 1.5.4[137]. Docking was performed within a rectangular box of 60Å *x* 90Å *x* 60Å dimensions. The center of the grid was placed at the center of mass of the original peptide in its bound conformation in crystal structure 4IB5. Fifty independent docking runs were carried out for each peptide starting from random positions. For comparison, we also performed redocking of the native peptide to the respective crystal structure of the protein.

A separate set of docking runs was performed where we included one particular water molecule (WAT554) that is hydrogen bonded to the carbonyl oxygen atoms of Tyr188 of chain D and Pro108 of chain A in the crystal structure. As validation of this strategy, the program Dowser [61] was used that identifies putative internal hydration sites inside proteins. That program first identifies internal cavities in a protein structure and then assesses the hydrophilicity of these cavities in terms of the interaction energy of a water molecule with the surrounding atoms. For our system, Dowser identified only a single favorable water position. This position is near the binding cavity and is only 0.5 Å away from the position of water molecule WAT554 in chain A of the crystal structure. Thus, we kept this water molecule rigid and treated it as part of the receptor in our docking experiments. Since the crystal water did not bear the hydrogen atoms, we used the water molecule positioned by Dowser. Including the water molecule improved the results slightly for all peptides except the original one.

6.4.3. Results and Discussion

Table 6.1 lists the results from the molecular docking experiments. As seen in the table, the best scores and their corresponding average scores are very close which can be considered as a sign of convergence. Except for peptides 3,4-Cl_2-Phe and 3-CF_3-Phe, all other derivatives of the Pc peptide showed more favorable binding affinity compared to the Pc peptide itself. Our colleagues further analyzed the cyclic peptides including amino acid 3-iodo-L-Phe and 3-chloro-L-Phe that became prominent in the molecular docking experiments and due to their commercial availability. The compound 3-iodo-L-Phe was of particular

interest since iodine substituents are known to form stronger halogen bonds than chlorine and bromine atoms, due to the larger atomic radius and the high polarizability. Using isothermal calorimetry, our colleagues showed that the 3-iodo-L-Phe containing peptide has a more favorable binding affinity than the peptide that contains 3-chloro-L-Phe. Resolving such fine details is evidently beyond the ability of the molecular docking procedure followed here[150].

Table 6.1: Calculated binding energies obtained by docking of Pc and Pc derivatives into the ligand free-free crystal structure of CK2α.

	Best score (kcal/mol)		Average Score (kcal/mol)*	
	-W554	+W554	-W554	+W554
Phe (Pc)	-5.75	-5.74	-5.74	-5.70
3-CN-Phe	-6.15	-6.36	-6.01	-6.34
3,4-Cl_2-Phe	-4.24	-4.64	-4.05	-4.55
3-Cl-Phe	-6.17	-6.34	-6.16	-6.32
3-F-Phe	-5.82	-5.94	-5.81	-5.89
3-I-Phe	-5.90	-6.39	-5.87	-6.36
3-OH-Phe	-6.01	-6.19	-5.99	-6.10
3-CF_3-Phe	-5.53	-5.77	-5.41	-5.61

* Average score over 50 independent runs.

7. Conclusions and Outlook

In this thesis, we used different molecular modeling techniques to study the mechanisms of molecular association reactions and to support experimental analyses. Using molecular dynamics simulations we studied structural and energetics aspects of specific (Chapter 3) and non-specific protein-protein complexes (Chapter 4) formed by hydrophilic proteins. Utilizing the alchemical free energy perturbation calculations we quantified the allosteric coupling energy between two distant sites on phosphoinostide-dependent kinase-1 (Chapter 5). Lastly, we showed the power of molecular docking for modeling protein-peptide interaction.

Previous studies showed that protein-protein interfaces contain more charged and polar residues than nonpolar residues. Therefore, hydrophilic protein-protein interfaces constitute an important if not the major part of all protein-protein interfaces. So far, the driving force behind their association has remained poorly characterized in the literature. Using three well-studied protein-protein complexes, we probed the association of hydrophilic proteins while forming specific complexes. The studied complexes are Barnase-Barstar (BN-BS), Cytochrome *c* – Cytochrome *c* peroxidase (CC-CC), and the complex of the N-terminal domain of enzyme I with Histidine-containing Phosphocarrier (EIN-HPr). The one-dimensional free energy profiles of protein-protein association were obtained from umbrella sampling simulations. The standard free energies of binding computed from the one-dimensional free energy profiles are in overall good agreement with the experimental values. We decomposed the standard free energy of binding into three terms, which are the energy change due to the change in translational and rotational degrees of freedom of all species (the monomers and the complex), the energy change for bringing the two proteins from infinite separation to the final configuration of the complex in vacuum, and the solvent induced contribution. This revealed that the favorable electrostatic and Lennard Jones interactions between the protein pairs render the solvent-induced interactions repulsive. The interfacial water between the two proteins was characterized by means of density and orientational order parameter. The

orientational order parameter of confined water deviates to a small degree but noticeably from bulk values, especially at close separations of the proteins. At close separations of the complex partners, the water at the interfacial gap is found to be more dense compared to bulk water. This theoretical analysis illustrated that the bound states of these 3 hydrophilic protein systems is stabilized by quite similar physicochemical parameters. For the first time, we showed that the favorable electrostatic interactions extend to protein-protein distances of only about 1.5 nm. Beyond this separation, the free energy profile is flat.

For comparison, we have contrasted these findings with analyzing non-specific encounters formed by the same 3 protein pairs. This is intended to explore the situation in the crowded environment of cells where proteins frequently encounter other proteins in many possible orientations. Most of these encounters are short-lived because the two binding patches do not match in terms of physico-chemical properties. Not all of these encounters are undesirable and artifacts of the crowded environment of the living cell. Instead, nonspecific interactions contribute critically to the formation of specific complexes. Besides accelerating the binding kinetics, nonspecific complexes also play a role in protein function as alternative binding modes. First we carried out unbiased MD simulations of the BN–BS, CC-CCP and EIN-HPr systems started from separated positions to obtain a set of spontaneously forming nonspecific complexes. Afterwards, we applied a pre-filter to the formed complexes and picked two nonspecific complexes those with the longest lifetime and with the largest contact interface for detailed analyses. In comparison to their specific counterparts, non-specific encounters bear smaller interaction interfaces and are attracted by shorter-ranged direct interactions between the proteins. We found that the entropic contributions are almost the same as for the specific complexes. Therefore, the balance between direct interactions and the solvent-induced interactions determines the binding affinity, which is the main difference between nonspecific complexes and their specific counterparts.

Proteins populate several switchable states. If a triggering event that effectively switches the protein population from one state to the other is far away from the

active site of the protein, it is termed allostery. Allostery plays an important role in almost all processes in living cells and in drug discovery. Phosphoinostide-dependent kinase-1 (PDK1) is a "master" protein kinase in insulin and growth factor signaling. PDK1 has an allosteric site termed PIF-pocket, which is located at the N-terminal lobe of the kinase domain, where the substrate kinases dock their phosphorylated conserved hydrophobic motif. This docking event enables PDK1 to become fully active and phosphorylate its substrate. Several allosteric modulators of PDK1 have been developed. In order to quantify the allosteric effect, we calculated the allosteric coupling energy between the ATP binding pocket and the PIF-pocket of PDK1 as the difference of the two binding free energies computed in the presence and absence of the allosteric modulator PS182. For this, we designed a thermodynamic cycle. Then, alchemical free energy perturbation calculations were performed to compute the standard binding free energies. For the studied system, the main contribution to the allosteric coupling energy comes from the electrostatic interactions. The contribution due to Lennard-Jones interactions did not change notably based on the method used. Studying the structural and thermodynamic properties of charged and highly polar compounds requires particular care especially when the compounds are highly flexible. In such cases, the flexibility of the molecule causes large fluctuations in the electrostatic interactions, which eventually introduces large statistical errors.

The ER luminal binding immunoglobulin protein (BIP) and cytosolic Ca^{2+}-Calmodulin facilitate the gating of Sec61 from the open to the closed state. Using molecular docking, we analyzed the putative interactions of BIP with loop 7 of Sec61. The results obtained from molecular docking were in parallel with findings from peptide binding experiments performed by our collaborators.

Casein kinase II (CK2) is a highly conserved Ser/Thr kinase composed of two catalytic α-subunits (CK2α) and two regulatory β-subunits (CK2β). In another collaboration project, we used a cyclic peptide as a scaffold to design new CK2β competitive compounds that bind to CK2α using molecular docking. The cyclic peptides including amino acid 3-iodo-L-Phe and 3-chloro-L-Phe performed better than the scaffold peptide.

The largest part of the work documented in this thesis aimed at elucidating the mechanisms how hydrophilic proteins form specific and nonspecific complexes and how they differ from each other. Very extensive molecular dynamics calculations have been performed for that task and the topic has been approached from various aspects. The interfaces of both specific and nonspecific complexes studied here were of the same character. In future work, it would be interesting to apply the same simulation methodology to study formation of complexes formed by two hydrophobic proteins. Another interesting topic we have not covered here the reciprocity of allostery. We have quantified the allosteric effect at the active site. One can, in principle, also quantify the allosteric effect at the allosteric site. For this, one should pick a relatively simple system with apolar or lowly polar ligands.

8. Appendix

Table A.1. The PDK crystal structures used for principal component analysis. Only the structures with at least 310 amino acids were used.

PDB ID	Chain
3rcj	A
4rrv	A
4rqv	A
4rqk	A
4aw1	A
4aw0	A
4a07	A
4a06	A
3sc1	A
3rwq	A
3rwp	A
3qd4	A
3qd3	A
3qd0	A
3qcy	A
3qcx	A
3qcs	A
3qc4	A
3pwy	A
3orz	A
3orx	A
3nay	A
3nax	A
3iop	A
3ion	A
3hrf	A
3hrc	A
3h9o	A
2r7b	A
2biy	A
1uu8	A
1uu7	A
1uu3	A
1okz	A
1oky	A
3qcq	A
1h1w	A

9. References

1. Jackson, M.B., *Molecular and Cellular Biophysics*. 2006, New York: Cambridge University Press.
2. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels.* Nature Structural Biology, 1997. **4**(1): p. 10-19.
3. Tsai, C.J., et al., *Folding funnels, binding funnels, and protein function.* Protein Science, 1999. **8**(6): p. 1181-1190.
4. Bryngelson, J.D., et al., *FUNNELS, PATHWAYS, AND THE ENERGY LANDSCAPE OF PROTEIN-FOLDING - A SYNTHESIS.* Proteins-Structure Function and Genetics, 1995. **21**(3): p. 167-195.
5. Chandler, D., *Interfaces and the driving force of hydrophobic assembly.* Nature, 2005. **437**(7059): p. 640-647.
6. Gilson, M.K., et al., *The statistical-thermodynamic basis for computation of binding affinities: A critical review.* Biophysical Journal, 1997. **72**(3): p. 1047-1069.
7. General, I.J., *A Note on the Standard State's Binding Free Energy.* Journal of Chemical Theory and Computation, 2010. **6**(8): p. 2520-2524.
8. Hill, T.L., *Cooperativity Theory in Biochemistry*. 1 ed. Springer Series in Molecular Biology. 1985, New York: Springer-Verlag.
9. Hill, T.L., *THEORY OF PROTEIN SOLUTIONS .1.* Journal of Chemical Physics, 1955. **23**(4): p. 623-636.
10. Verlet, L., *COMPUTER EXPERIMENTS ON CLASSICAL FLUIDS .I. THERMODYNAMICAL PROPERTIES OF LENNARD-JONES MOLECULES.* Physical Review, 1967. **159**(1): p. 98-&.
11. Hockney, R.W., S.P. Goel, and J.W. Eastwood, *QUIET HIGH-RESOLUTION COMPUTER MODELS OF A PLASMA.* Journal of Computational Physics, 1974. **14**(2): p. 148-158.
12. Born, M. and R. Oppenheimer, *Zur Quantentheorie der Molekeln*. 1927, Annalen der Physik. p. 457-484.

13. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
14. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
15. Schuler, L.D., X. Daura, and W.F. Van Gunsteren, *An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase.* Journal of Computational Chemistry, 2001. **22**(11): p. 1205-1218.
16. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
17. Aqvist, J., C. Medina, and J.E. Samuelsson, *NEW METHOD FOR PREDICTING BINDING-AFFINITY IN COMPUTER-AIDED DRUG DESIGN.* Protein Engineering, 1994. **7**(3): p. 385-391.
18. Hansson, T., J. Marelius, and J. Aqvist, *Ligand binding affinity prediction by linear interaction energy methods.* Journal of Computer-Aided Molecular Design, 1998. **12**(1): p. 27-35.
19. Srinivasan, J., et al., *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices.* Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.
20. Gilson, M.K. and H.-X. Zhou, *Calculation of protein-ligand binding affinities.* Annual Review of Biophysics and Biomolecular Structure, 2007. **36**: p. 21-42.
21. Potter, M.J. and M.K. Gilson, *Coordinate systems and the calculation of molecular properties.* Journal of Physical Chemistry A, 2002. **106**(3): p. 563-566.
22. Tembe, B.L. and J.A. McCammon, *LIGAND RECEPTOR INTERACTIONS.* Computers & Chemistry, 1984. **8**(4): p. 281-283.
23. Hermans, J. and S. Shankar, *THE FREE-ENERGY OF XENON BINDING TO MYOGLOBIN FROM MOLECULAR-DYNAMICS SIMULATION.* Israel Journal of Chemistry, 1986. **27**(2): p. 225-227.

24. Jorgensen, W.L., et al., *EFFICIENT COMPUTATION OF ABSOLUTE FREE-ENERGIES OF BINDING BY COMPUTER-SIMULATIONS - APPLICATION TO THE METHANE DIMER IN WATER.* Journal of Chemical Physics, 1988. **89**(6): p. 3742-3746.
25. Zwanzig, R.W., *HIGH-TEMPERATURE EQUATION OF STATE BY A PERTURBATION METHOD .1. NONPOLAR GASES.* Journal of Chemical Physics, 1954. **22**(8): p. 1420-1426.
26. Jorge, M., et al., *Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration.* Journal of Chemical Theory and Computation, 2010. **6**(4): p. 1018-1027.
27. Shyu, C. and F.M. Ytreberg, *Reducing the Bias and Uncertainty of Free Energy Estimates by Using Regression to Fit Thermodynamic Integration Data.* Journal of Computational Chemistry, 2009. **30**(14): p. 2297-2304.
28. Bennett, C.H., *EFFICIENT ESTIMATION OF FREE-ENERGY DIFFERENCES FROM MONTE-CARLO DATA.* Journal of Computational Physics, 1976. **22**(2): p. 245-268.
29. Shirts, M.R., et al., *Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods.* Physical Review Letters, 2003. **91**(14).
30. Kirkwood, J.G., *Statistical mechanics of fluid mixtures*. 1935, The Journal of Chemical Physics. p. 300-313.
31. Roux, B., *THE CALCULATION OF THE POTENTIAL OF MEAN FORCE USING COMPUTER-SIMULATIONS.* Computer Physics Communications, 1995. **91**(1-3): p. 275-282.
32. Torrie, G.M. and J.P. Valleau, *NON-PHYSICAL SAMPLING DISTRIBUTIONS IN MONTE-CARLO FREE-ENERGY ESTIMATION - UMBRELLA SAMPLING.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
33. Kumar, S., et al., *The Weighted Histogram Analyses Method for Free-Energy Calculations on Biomolecules .1. The Method.* Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.
34. Ferrenberg, A.M. and R.H. Swendsen, *OPTIMIZED MONTE-CARLO DATA-ANALYSIS.* Physical Review Letters, 1989. **63**(12): p. 1195-1198.

35. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states.* Journal of Chemical Physics, 2008. **129**(12).
36. Tan, Z.Q., *On a likelihood approach for Monte Carlo integration.* Journal of the American Statistical Association, 2004. **99**(468): p. 1027-1036.
37. Waksman, G., *Proteomics and Protein-Protein Interactions: Biology, Chemistry, Bioinformatics, and Drug Design*. ProteinReviews, ed. M.Z. Atassi. 2005, New York.
38. Sheinerman, F.B., R. Norel, and B. Honig, *Electrostatic aspects of protein-protein interactions.* Current Opinion in Structural Biology, 2000. **10**(2): p. 153-159.
39. Ansari, S. and V. Helms, *Statistical analysis of predominantly transient protein-protein interfaces.* Proteins-Structure Function and Bioinformatics, 2005. **61**(2): p. 344-355.
40. Larsen, T.A., A.J. Olson, and D.S. Goodsell, *Morphology of protein-protein interfaces.* Structure with Folding & Design, 1998. **6**(4): p. 421-427.
41. Bryngelson, J.D., et al., *Funnels, Pathways, and the Energy Landscape of Protein Folding - A Synthesis.* Proteins-Structure Function and Genetics, 1995. **21**(3): p. 167-195.
42. Woo, H.J. and B. Roux, *Calculation of absolute protein-ligand binding free energy from computer simulations.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(19): p. 6825-6830.
43. Kaestner, J., *Umbrella sampling.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2011. **1**(6): p. 932-942.
44. McLain, S.E., et al., *Charge-Based Interactions between Peptides Observed as the Dominant Force for Association in Aqueous Solution.* Angewandte Chemie-International Edition, 2008. **47**(47): p. 9059-9062.
45. Ben-Naim, A., *On the driving forces for protein-protein association.* Journal of Chemical Physics, 2006. **125**(2): p. 24901–24911.
46. Hua, L., R. Zangi, and B.J. Berne, *Hydrophobic Interactions and Dewetting between Plates with Hydrophobic and Hydrophilic Domains.* Journal of Physical Chemistry C, 2009. **113**(13): p. 5244-5253.

47. Ahmad, M., W. Gu, and V. Helms, *Mechanism of fast peptide recognition by SH3 domains.* Angewandte Chemie-International Edition, 2008. **47**(40): p. 7626-7630.
48. Ahmad, M., et al., *Adhesive water networks facilitate binding of protein interfaces.* Nature Communications, 2011. **2**: p. -.
49. Buckle, A.M., G. Schreiber, and A.R. Fersht, *PROTEIN-PROTEIN RECOGNITION - CRYSTAL STRUCTURAL-ANALYSIS OF A BARNASE BARSTAR COMPLEX AT 2.0-ANGSTROM RESOLUTION.* Biochemistry, 1994. **33**(30): p. 8878-8889.
50. Gabdoulline, R.R. and R.C. Wade, *Simulation of the diffusional association of Barnase and Barstar.* Biophysical Journal, 1997. **72**(5): p. 1917-1929.
51. Spaar, A. and V. Helms, *Free energy landscape of protein-protein encounter resulting from Brownian dynamics simulations of Barnase : Barstar.* Journal of Chemical Theory and Computation, 2005. **1**(4): p. 723-736.
52. Vijayakumar, M., et al., *Electrostatic enhancement of diffusion-controlled protein-protein association: Comparison of theory and experiment on barnase and barstar.* Journal of Molecular Biology, 1998. **278**(5): p. 1015-1024.
53. Gumbart, J.C., B. Roux, and C. Chipot, *Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles.* Journal of Chemical Theory and Computation, 2013. **9**(8): p. 3789-3798.
54. Pelletier, H. and J. Kraut, *Crystal-Structure of a Complex Between Electron-Transfer Partners, Cytochrome-c Peroxidase and Cytochrome-c.* Science, 1992. **258**(5089): p. 1748-1755.
55. Northrup, S.H., J.O. Boles, and J.C.L. Reynolds, *Brownian Dynamics of Cytochrome-c and Cytochrome-c Peroxidase Association.* Science, 1988. **241**(4861): p. 67-70.
56. Gabdoulline, R.R. and R.C. Wade, *Protein-protein association: Investigation of factors influencing association rates by Brownian dynamics simulations.* Journal of Molecular Biology, 2001. **306**(5): p. 1139-1155.
57. Garrett, D.S., et al., *Solution structure of the 40,000 M-r phosphoryl transfer complex between the N-terminal domain of enzyme I and HPr.* Nature Structural Biology, 1999. **6**(2): p. 166-173.

58. Kim, Y.C., et al., *Replica exchange simulations of transient encounter complexes in protein-protein association.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(35): p. 12855-12860.
59. Bernstein, F.C., et al., *Protein Data Bank - Computer-Based Archival File For Macromolecular Structures.* Journal of Molecular Biology, 1977. **112**(3): p. 535-542.
60. Li, H., A.D. Robertson, and J.H. Jensen, *Very fast empirical prediction and rationalization of protein pK(a) values.* Proteins-Structure Function and Bioinformatics, 2005. **61**(4): p. 704-721.
61. Zhang, L. and J. Hermans, *Hydrophilicity of cavities in proteins.* Proteins-Structure Function and Genetics, 1996. **24**(4): p. 433-438.
62. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins-Structure Function and Bioinformatics, 2010. **78**(8): p. 1950-1958.
63. Essmann, U., et al., *A SMOOTH PARTICLE MESH EWALD METHOD.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
64. Jorgensen, W., *Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water.* J. Am. Chem. Soc, 1981. **103**(2): p. 335-340.
65. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
66. Worrall, J.A.R., et al., *Interaction of yeast iso-1-cytochrome c with cytochrome c peroxidase investigated by N-15,H-1 heteronuclear NMR spectroscopy.* Biochemistry, 2001. **40**(24): p. 7069-7076.
67. Frisch, M.J., G.W. Trucks, and H.B. Schlegel, *Gaussian 03, Revision C.02*. 2004, Gaussian, Inc. Wallingford CT.
68. Cieplak, P., et al., *Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, RNA and Proteins.* Journal of Computational Chemistry, 1995. **16**(11): p. 1357-1377.

69. Shahrokh, K., et al., *Quantum Mechanically Derived AMBER-Compatible Heme Parameters for Various States of the Cytochrome P450 Catalytic Cycle.* Journal of Computational Chemistry, 2012. **33**(2): p. 119-133.
70. Berendsen, H.J.C., et al., *MOLECULAR-DYNAMICS WITH COUPLING TO AN EXTERNAL BATH.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
71. Hess, B., *P-LINCS: A parallel linear constraint solver for molecular simulation.* Journal of Chemical Theory and Computation, 2008. **4**(1): p. 116-122.
72. Nose, S., *A UNIFIED FORMULATION OF THE CONSTANT TEMPERATURE MOLECULAR-DYNAMICS METHODS.* Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
73. Hoover, W.G., *CANONICAL DYNAMICS - EQUILIBRIUM PHASE-SPACE DISTRIBUTIONS.* Physical Review A, 1985. **31**(3): p. 1695-1697.
74. Parrinello, M. and A. Rahman, *Polymorphic Transitions in Single-Crystals - A New Molecular-Dynamics Method.* Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
75. Hub, J.S., B.L. de Groot, and D. van der Spoel, *g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates.* Journal of Chemical Theory and Computation, 2010. **6**(12): p. 3713-3720.
76. Doudou, S., N.A. Burton, and R.H. Henchman, *Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force.* Journal of Chemical Theory and Computation, 2009. **5**(4): p. 909-918.
77. Lazaridis, T., A. Masunov, and F. Gandolfo, *Contributions to the binding free energy of ligands to avidin and streptavidin.* Proteins-Structure Function and Genetics, 2002. **47**(2): p. 194-208.
78. Irudayam, S.J. and R.H. Henchman, *Entropic Cost of Protein-Ligand Binding and Its Dependence on the Entropy in Solution.* Journal of Physical Chemistry B, 2009. **113**(17): p. 5871-5884.
79. Peter, W., A. Sam, and H. Volkhard, *ABC (Analysing Biomolecular Contacts)-database*. 2007: Journal of Integrative Bioinformatics.
80. Kuntz, I.D., Jr. and W. Kauzmann, *Hydration of proteins and polypeptides.* Advances in protein chemistry, 1974. **28**: p. 239-345.

81. Squire, P.G. and M.E. Himmel, *Hydrodynamics and Protein Hydration.* Archives of Biochemistry and Biophysics, 1979. **196**(1): p. 165-177.
82. Gekko, K. and H. Noguchi, *Compressibility of Globular-Proteins in Water at 25-Degrees-C.* Journal of Physical Chemistry, 1979. **83**(21): p. 2706-2714.
83. Chau, P.L. and A.J. Hardwick, *A new order parameter for tetrahedral configurations.* Molecular Physics, 1998. **93**(3): p. 511-518.
84. Errington, J.R. and P.G. Debenedetti, *Relationship between structural order and the anomalies of liquid water.* Nature, 2001. **409**(6818): p. 318-321.
85. Baker, N.A., et al., *Electrostatics of nanosystems: Application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10037-10041.
86. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.
87. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(1): p. 13-20.
88. Schreiber, G. and A.R. Fersht, *Interaction of Barnase with Its Polypeptide Inhibitor Barstar Studied by Protein Engineering.* Biochemistry, 1993. **32**(19): p. 5145-5150.
89. Mei, H.K., et al., *Control of formation and dissociation of the high-affinity complex between cytochrome c and cytochrome c peroxidase by ionic strength and the low-affinity binding site.* Biochemistry, 1996. **35**(49): p. 15800-15806.
90. Ginsburg, A., et al., *Conformational stability changes of the amino terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate : sugar phosphotransferase system produced by substituting alanine or glutamate for the active-site histidine 189: Implications for phosphorylation effects.* Protein Science, 2000. **9**(6): p. 1085-1094.
91. Wang, L., et al., *Downhill Binding Energy Surface of the Barnase-Barstar Complex.* Biopolymers, 2010. **93**(11): p. 977-985.
92. Neumann, J. and K.-E. Gottschalk, *The Effect of Different Force Applications on the Protein-Protein Complex Barnase-Barstar.* Biophysical Journal, 2009. **97**(6): p. 1687-1699.

93. Buch, I., S. Kashif Sadiq, and G. De Fabritiis, *Optimized Potential of Mean Force Calculations for Standard Binding Free Energies.* Journal of Chemical Theory and Computation, 2011. **7**(6): p. 1765-1772.
94. Yu, Y.B., P.L. Privalov, and R.S. Hodges, *Contribution of translational and rotational motions to molecular association in aqueous solution.* Biophysical Journal, 2001. **81**(3): p. 1632-1642.
95. Zhou, H.-X. and M.K. Gilson, *Theory of Free Energy and Entropy in Noncovalent Binding.* Chemical Reviews, 2009. **109**(9): p. 4092-4107.
96. Ben-Naim, A., *On the driving forces for protein-protein association.* Journal of Chemical Physics, 2006. **125**(2).
97. Ball, P., *Water as an active constituent in cell biology.* Chemical Reviews, 2008. **108**(1): p. 74-108.
98. Rasaiah, J.C., S. Garde, and G. Hummer, *Water in nonpolar confinement: From nanotubes to proteins and beyond.* Annual Review of Physical Chemistry, 2008. **59**: p. 713-740.
99. Yan, Z., et al., *Structure of the first- and second-neighbor shells of simulated water: Quantitative relation to translational and orientational order.* Physical Review E, 2007. **76**(5).
100. Danielewicz-Ferchmin, I. and A.R. Ferchmin, *Phase diagram of electrostricted H2O at surfaces of electrodes at 273-373 K: Electric critical point of water.* Chemphyschem, 2005. **6**(8): p. 1499-1509.
101. Danielewicz-Ferchmin, I. and A.R. Ferchmin, *Water at ions, biomolecules and charged surfaces.* Physics and Chemistry of Liquids, 2004. **42**(1): p. 1-36.
102. Fulton, A.B., *HOW CROWDED IS THE CYTOPLASM.* Cell, 1982. **30**(2): p. 345-347.
103. Ellis, R.J., *Macromolecular crowding: obvious but underappreciated.* Trends in Biochemical Sciences, 2001. **26**(10): p. 597-604.
104. Ubbink, M., *The courtship of proteins: Understanding the encounter complex.* Febs Letters, 2009. **583**(7): p. 1060-1066.
105. Iwahara, J. and G.M. Clore, *Detecting transient intermediates in macromolecular binding by paramagnetic NMR.* Nature, 2006. **440**(7088): p. 1227-1230.

106. Ulucan Ozlem, J.T., and Helms Volkhard, *Energetics of Hydrophilic Protein–Protein Association and the Role of Water.* Journal of Chemical Theory and Computation, 2014. **10**(8): p. 3512–3524.
107. Hoefling, M. and K.E. Gottschalk, *Barnase-Barstar: From first encounter to final complex.* Journal of Structural Biology, 2010. **171**(1): p. 52-63.
108. Schreiber, G. and A.R. Fersht, *Rapid, electrostatically assisted association of proteins.* Nature Structural Biology, 1996. **3**(5): p. 427-431.
109. Spaar, A., et al., *Diffusional encounter of barnase and barstar.* Biophysical Journal, 2006. **90**(6): p. 1913-1924.
110. Volkov, A.N., et al., *Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(50): p. 18945-18950.
111. Tang, C., J. Iwahara, and G.M. Clore, *Visualization of transient encounter complexes in protein-protein association.* Nature, 2006. **444**(7117): p. 383-386.
112. Tsai, C.-J. and R. Nussinov, *A Unified View of "How Allostery Works".* Plos Computational Biology, 2014. **10**(2).
113. P., K.T., *Allosteric Drug Effects* in *Pharmacology in Drug Discovery: Understanding Drug Response*. 2011, Academic Press: Chapel Hill.
114. Ricci, F., et al., *Rational Design of Allosteric Inhibitors and Activators Using the Population-Shift Model: In Vitro Validation and Application to an Artificial Biosensor.* Journal of the American Chemical Society, 2012. **134**(37): p. 15177-15180.
115. Kostenis, E. and K. Mohr, *Two-point kinetic experiments to quantify allosteric effects on radioligand dissociation.* Trends in Pharmacological Sciences, 1996. **17**(8): p. 280-283.
116. Huwiler, K.G., et al., *A Fluorescence Anisotropy Assay for the Muscarinic M1 G-protein-Coupled Receptor.* Assay and Drug Development Technologies, 2010. **8**(3): p. 356-366.
117. Mora, A., et al., *PDK1, the master regulator of AGC kinase signal transduction.* Seminars in Cell & Developmental Biology, 2004. **15**(2): p. 161-170.

118. Biondi, R.M., et al., *High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site.* The EMBO journal, 2002. **21**(16): p. 4219-4228.
119. Komander, D., et al., *Structural insights into the regulation of PDK1 by phosphoinositides and inositol phosphates.* The EMBO journal, 2004. **23**(20): p. 3918-3928.
120. Engel, M., et al., *Allosteric activation of the protein kinase PDK1 with low molecular weight compounds.* The EMBO journal, 2006. **25**(23): p. 5469-5480.
121. Stroba, A., et al., *3, 5-Diphenylpent-2-enoic Acids as Allosteric Activators of the Protein Kinase PDK1: Structure- Activity Relationships and Thermodynamic Characterization of Binding as Paradigms for PIF-Binding Pocket-Targeting Compounds†† PDB code of 2Z with PDK1: 3HRF.* Journal of medicinal chemistry, 2009. **52**(15): p. 4683-4693.
122. Busschots, K., et al., *Substrate-Selective Inhibition of Protein Kinase PDK1 by Small Compounds that Bind to the PIF-Pocket Allosteric Docking Site.* Chemistry & Biology, 2012. **19**(9): p. 1152-1163.
123. Hindie, V., et al., *Structure and allosteric effects of low-molecular-weight activators on the protein kinase PDK1.* Nature chemical biology, 2009. **5**(10): p. 758-764.
124. Mobley, D.L., J.D. Chodera, and K.A. Dill, *On the use of orientational restraints and symmetry corrections in alchemical free energy calculations.* Journal of Chemical Physics, 2006. **125**(8).
125. Wang, J., et al., *Development and testing of a general amber force field.* Journal of computational chemistry, 2004. **25**(9): p. 1157-1174.
126. Boresch, S., et al., *Absolute binding free energies: A quantitative approach for their calculation.* Journal of Physical Chemistry B, 2003. **107**(35): p. 9535-9551.
127. Hornak, V. and C. Simmerling, *Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations.* Journal of Molecular Graphics & Modelling, 2004. **22**(5): p. 405-413.
128. Shirts, M.R. and V.S. Pande, *Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio,*

and thermodynamic integration. Journal of Chemical Physics, 2005. **122**(14).

129. Biondi, R.M., et al., *High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site.* Embo Journal, 2002. **21**(16): p. 4219-4228.
130. Amadei, A., A.B.M. Linssen, and H.J.C. Berendsen, *ESSENTIAL DYNAMICS OF PROTEINS.* Proteins-Structure Function and Genetics, 1993. **17**(4): p. 412-425.
131. Johnson, D.A., et al., *Dynamics of cAMP-dependent protein kinase.* Chemical Reviews, 2001. **101**(8): p. 2243-2270.
132. Leach, A.R., *Molecular Modelling: Principles and Applications*. 2 ed. 2001: Prentice Hall.
133. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: Methods and applications.* Nature Reviews Drug Discovery, 2004. **3**(11): p. 935-949.
134. Kuntz, I.D., et al., *A GEOMETRIC APPROACH TO MACROMOLECULE-LIGAND INTERACTIONS.* Journal of Molecular Biology, 1982. **161**(2): p. 269-288.
135. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
136. Solis, F.J. and R.J.B. Wets, *MINIMIZATION BY RANDOM SEARCH TECHNIQUES.* Mathematics of Operations Research, 1981. **6**(1): p. 19-30.
137. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility.* Journal of Computational Chemistry, 2009. **30**(16): p. 2785-2791.
138. Wesson, L. and D. Eisenberg, *ATOMIC SOLVATION PARAMETERS APPLIED TO MOLECULAR-DYNAMICS OF PROTEINS IN SOLUTION.* Protein Science, 1992. **1**(2): p. 227-235.
139. Dierks, T., et al., *A microsomal ATP-binding protein involved in efficient protein transport into the mammalian endoplasmic reticulum.* Embo Journal, 1996. **15**(24): p. 6931-6942.

140. Hamman, B.D., et al., *The aqueous pore through the translocon has a diameter of 40-60 angstrom during cotranslational protein translocation at the ER membrane.* Cell, 1997. **89**(4): p. 535-544.
141. Alder, N.N., et al., *The molecular mechanisms underlying BiP-mediated gating of the Sec61 translocon of the endoplasmic reticulum.* Journal of Cell Biology, 2005. **168**(3): p. 389-399.
142. Sali, A. and T.L. Blundell, *COMPARATIVE PROTEIN MODELING BY SATISFACTION OF SPATIAL RESTRAINTS.* Journal of Molecular Biology, 1993. **234**(3): p. 779-815.
143. Schaeuble, N., et al., *BiP-mediated closing of the Sec61 channel limits Ca2+ leakage from the ER.* Embo Journal, 2012. **31**(15): p. 3282-3296.
144. Cozza, G., F. Meggio, and S. Moro, *The Dark Side of Protein Kinase CK2 Inhibition.* Current Medicinal Chemistry, 2011. **18**(19): p. 2867-2884.
145. Laudet, B., et al., *Structure-based design of small peptide inhibitors of protein kinase CK2 subunit interaction.* Biochemical Journal, 2007. **408**: p. 363-373.
146. Laudet, B., et al., *Identification of chemical inhibitors of protein-kinase CK2 subunit interaction.* Molecular and Cellular Biochemistry, 2008. **316**(1-2): p. 63-69.
147. Raaf, J., et al., *First Structure of Protein Kinase CK2 Catalytic Subunit with an Effective CK2 beta-Competitive Ligand.* Acs Chemical Biology, 2013. **8**(5): p. 901-907.
148. Case, D., et al., *AMBER 10.* University of California, San Francisco, 2008.
149. *Discovery Studio Modeling Environment, Release 2.5*. 2009: *San Diego: Accelrys Software Inc*.
150. Hochscherf, J., et al., *Development of a high-throughput screening-compatible assay to identify inhibitors of the CK2a/CK2b interaction.* Analytical Biochemistry, 2014. **468**(2015): p. 4-14.