Computer Simulation of Biomolecular Solvation, Recognition and Proton Transfer Equilibria

Dissertation

zur Erlangung des Grades des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

vorgelegt von

Wei Gu

Saarbrücken, 2007

谨以此文献给我的父母,妻子和所有亲人

Acknowledgement

First and most, I would like to show my appreciation to my supervisor, Prof. Dr. Volkhard Helms for offering me the opportunity to work in the *Group of Computational Biology* at the University of Saarland. Your profound knowledge and experiences in computational biology and computer simulation and your grateful encouragement have been of great help throughout the past four years. Many thanks for your fruitful discussions, your correction of all my manuscripts and patiently improving my English. I appreciate you a lot for giving me so many opportunities to attend interesting conferences, workshops and seminars. I thank you for helping me to overcome many bureaucratic hurdles with visa extension and other administration issues.

I would also like to thank my former supervisors Prof. Haiyan Liu and Prof. Yunyu Shi at the University of Sciences and Technology of China for leading me to the field of computer simulation. Thank you for your valuable instructions at the early stage of my research career and your kind recommendations for my further study.

I want to thank all my colleagues in Saarbrücken and Frankfurt: Dr. Michael Hutter for many useful advices in quantum chemistry calculation and careful readings of many of my manuscripts, Dr. Tihamér Geyer for all kinds of helps with computer systems and physical knowledge, Dr. Böckmann for useful discussions on MD simulations, Tomaso, Elena(s), René and Mazen for working together with me, Yungki for taking all the hard lectures and finishing all the tough homework with me, Sam, Christian, Alex(es) and Saurabh for many fun times together, Peter and Sikander for helping me with my German and English, Gautam for delicious Indian food, Daniele for many useful Italian words, Denitsa and Susanne for daily talk and coffee. I also thank Shirley, Ling, Ines, Babara, Siti, Beate, Simon from the group of Dr. Helms and Dr. Böckmann. I would like to thank our secretary Kerstin for helping me with a lot of paper works.

I thank my senior friend Jiang for teaching me from the very beginning of my graduate study. I thank Song, Hao, Haibo, Li, Qianqian, Peng, Chao, Jun(W), and Yingyu from USTC for your friendship and love.

My appreciations to Yisuo, Qian(L), Jiandong, Tao, Junda, Guanfeng, Junfeng, Degang, Kuangyu, Hai and other friends from *Kentanhui* for your enlightening advices and discussion and taking care of my daily life issues outside the research domain. Thanks also go to Dan, Feng, Zizhuo, Wei(D), Hanglin, Xiaohung, Xiwen, Rui, Hongbo, Dongmei and many other friends in Germany for various kinds of help.

I want to specially thank my wife Bin for your love, your support to my work, and sharing your life with me. 多少言语也无法表达我对父母的爱和感谢，您们无私的爱和对我的抚育培养是我今天一切的源泉。我深深的怀念我的母亲，未能及时报答您的爱是我今生最大的遗憾，您伟大的母爱和音容笑貌将永远在我的心中陪伴我。我也要感谢我的其他亲人在我各个时期对我的关怀，支持和帮助。

I thank the Volkswagen Foundation for the financial support of my study and research.

Contents

Abstract

In the past decays, computer simulations technique became a powerful tool in biophysics, material sciences as well as in energy chemistry. Computer simulations, especially molecular dynamics (MD) simulation, can provide the ultimate detail concerning individual particle motions as a function of time, therefore, in today's research, computer simulations are used to address many specific questions and details that are of interest for biomolecular functions. In this thesis, we studied several biological/chemical systems using standard and variant molecular dynamics simulation techniques. In the simulations of polyproline peptides interacting with their binding domains, we identified the solvent conformations of the unbounded peptides: the formation of a PPII helix of the peptide is not induced by the binding processes alone. Peptide docking and subsequent MD simulations of the G8X mutants identified an alternative binding mode, where a shift in register for the interacting prolines was observed. In the calculation of the solvation free energies of peptides of various lengths using the multiconfiguration thermodynamic integration and the Generalized born surface area implicit solvent model, non-additivity of the solvation free energies is found by both methodologies for peptides shorter than 5 residues. This non-additivity shows that the design of simplified models, where peptides and proteins are composed of residue-beads and interactions are modeled additively, appears challenging. We also investigated the dynamic protonation equilibria of acetic acid (AC^-/ACH) and 4-methylimidazole ($4MIH^+$/4MI) in aqueous solution with nearby proton accepting groups using the Q-HOP MD methods. In the simulations of acetic acid, we observed two different regimes of proton transfer: Extended phases of frequent proton swapping between acetic acid and nearby water were separated by phases where the proton freely diffuses in the simulation box until it is captured again by acetic acid. In the studied of 4-methylimidazole in aqueous solution and with nearby proton accepting groups, qualitatively different protonation behavior of 4-methylimidazole compared to that of acetic acid was found: On one hand, $4MIH^+$ has a high tendency to keep a proton once it is bound. On the other hand, 4MI has a relatively small proton capture radius, making it very hard to attract protons from long distances. Protonated acetic acid can easily share the proton with close titratable groups even if the acceptor group has a low pK_a. Moreover, AC^- has a large proton capture radius, making it a perfect proton "capturer".

Abstrakt

In den letzten Jahrzehnten entwickelten sich Computersimulationen zu einem leistungsstarken Instrument in den Bereichen Biophysik, Materialwissenschaft und Chemie. Insbesondere Moleküldynamik- (MD) Simulationen können detaillierte Informationen über individuelle Partikelbewegungen als Funktion über die Zeit liefern, weswegen in der heutigen Forschung solche Computersimulationen zur Klärung zahlreicher spezifischer Fragen und Details eingesetzt werden, die für biomolekulare Funktionen von Bedeutung sind. In dieser Arbeit untersuchte ich verschiedene biologische und chemische Systeme unter Verwendung von Standard- und alternativen Moleküldynamik-Simulationen. In Simulationen von Polyprolin-Peptiden, die mit ihren Bindungsdomänen interagieren, identifizierte ich die Solvenskonformation von ungebundenen Peptiden. Dabei zeigte sich, dass die Bildung einer PPII-Helix des Peptids nicht ausschließlich durch den Bindungsprozeß zustande kommt. Mit Hilfe von Peptid-Docking und anschließenden MD-Simulationen der G8X Mutante konnte ein alternativer Bindungsmodus aufgespürt werden, bei dem eine positionelle Verschiebung der an der Interaktion beteiligten Proline beobachtet wurde. Bei der Berechnung von freien Solvatationsenergien von Peptiden unterschiedlicher Länge mit Hilfe von „multiconfiguration thermodynamic integration“ und dem „generalized born surface area implicit solvent model“, wurde eine Nichtadditivität der freien Solvatationsenergien von beiden Methoden für Peptide mit weniger als fünf Residuen festgestellt. Diese Nicht-Additivität zeigt, dass das Design von vereinfachten Modellen, in denen Peptide und Proteine aus „residue beads“ bestehen und Interaktionen additiv modelliert werden, eine Herausforderung darstellt. Ich untersuchte ferner das dynamische Protonierungsgleichgewicht von Essigsäure (AC^-/ACH) und 4-Methylimidazol ($4MIH^+$/4MI) in wäßriger Lösung mit benachbarten Protonen-Akzeptorgruppen unter Verwendung der Q-HOP MD-Methoden. In den Simulationen mit Essigsäure beobachtete ich zwei verschiedene Systeme von Protontransfers. Ausgedehnte Phasen von häufigen Protonensprüngen zwischen Essigsäure und benachbarten Wassermolekülen können von Phasen unterschieden werden, in denen sich das Proton frei in der Simulationsbox bewegt, bis es wieder von einem Essigsäuremolekül eingefangen wird. Während der Untersuchung von 4-Methylimidazol in wäßriger Lösung mit benachbarten Protonakzeptorgruppen wurde eine qualitativ unterschiedliche Protonierung von 4-Methylimidazol im Vergleich zur Essigsäure beobachtet. Zum einen wies $4MIH^+$ eine starke Tendenz dazu auf, an dem Proton festzuhalten, sobald es gebunden war, andererseits kann 4MI nur innerhalb eines relativ kleinen Radius Protonen einfangen, so daß es ihm sehr schwer fällt, weiter entfernte Protonen anzuziehen. Protonierte Essigsäure kann das Proton leicht an nahe gelegene titrierbare Gruppen abgeben, selbst wenn der Akzeptor nur einen geringen pK_a-Wert hat. Darüber hinaus verfügt AC^- über eine weiten Protoneneinfang-Radius, der es zu einem perfekten „Protonenfänger“ macht.

Zusammenfassung

In dieser Arbeit untersuchten wir verschiedene biologische und chemische Systeme mit Hilfe von Standard- und speziellen MD-Simulationen.

Untersuchung der Interaktion zwischen prolinreichen Peptiden und ihren Adapterdomänen

Die Erkennung von prolinreichen Sequenzen spielt eine wichtige Rolle für das Entstehen von Multiprotein-Komplexen im Verlauf der Signaltransduktion von Eukaryonten und wird von einer Reihe von Proteinfaltungen vermittelt, die untereinander charakteristische Merkmale aufweisen.

Mit Hilfe von Molecular-Modelling und MD-Simulationen untersuchten wir die Solvenskonformation sowohl vom Wildtyp als auch von Mutanten von Polyprolin Peptiden, die an die GYF-Domäne binden. Wir fanden heraus, daß die Peptide selbst in Abwesenheit der GYF-Domäne PPII-Helix-Konformationen ausbildeten. Diese Ergebnisse stehen in Einklang mit kürzlich veröffentlichten experimentellen und theoretischen Studien zu prolinhaltigen und prolinfreien Polypeptiden und geben einen Hinweis darauf, daß die Bildung einer PPII Helix nicht durch den Bindungsprozeß allein begünstigt wird. Auf der Grundlage früherer Erkenntnisse aus NMR Studien der GYF-Domäne-Ligand Interaktion und Simulationen von Wildtyp- und mutierten Komplexen, modellierten wir den allgemeinen Bindungsmodus von Polyprolin-Peptiden der GYF-Domäne. Die hydrophoben Interaktionen zwischen den Peptidresiduen Pro6 und Pro7 und der Bindungstasche, die elektrostatische Anziehung zwischen den Peptidresiduen Arg3 und Arg10 und die Domänenresiduen Glu31 und Glu9 spielen wichtige Rollen bei der Bindung. Peptid-Docking und anschließende MD-Simulationen der G8X Mutanten brachten einen alternativen Bindungsmodus zum Vorschein, bei dem eine positionelle Verschiebung der interagierenden Proline zu beobachten war. Diese Ergebnisse stimmen qualitativ gut mit NMR chemical-shift-mapping Experimenten überein und geben eine Hinweis darauf, dass dynamische Proline für die Erkennung von prolinreichen Sequenzen von Bedeutung sind. Möglicherweise setzen solch gleitende Bewegungen entlang der prolinreichen Sequenzen die entropische Bestrafung der Bindung herab, zum anderen wird ein gewisser Grad an Spezifizität aufrecht erhalten.

Mit Hilfe von Peptiddocking und MD-Simulationen untersuchten wir die Bindung eines linearen Peptidmotifs an Cyclophilin A. Aus Substitutionsanalysen (Phage display) wurde der lineare Sequenzerkennungscode für CypA und dem Consensus Motif FGPXLp identifiziert (durchgeführt von mit uns kooperierenden Experimentalisten). Die modellierte Komplexstruktur stimmt sehr gut mit den Ergebnissen aus Phage-display Experimenten überein und liefert eine Erklärung für das spezifische Bindungsmotif hinsichtlich struktureller Gesichtspunkte und hinsichtlich der Interaktion.

Freie Solvatationsenergien von Peptiden unterschiedlicher Länge wurden mit Hilfe von MCTI und GBSA generiert. Für eine Residuenzahl von fünf oder mehr stehen die Ergebnisse recht gut im Einklang. Diese Tatsache bestärkt uns in unserem Vorhaben, ΔG_{hydr} für Peptide mit bis zu neun Residuen über MCTI Kalkulationen zu

berechnen. MCTI und GBSA zeigen jedoch für kurze Helices erhebliche Unterschiede auf, wobei MCTI eher akkurate Ergebnisse aufweisen sollte. Daher ist es wichtig die molekularen Details der Backbone-Hydratation zu berücksichtigen. Nicht-Additivität wurde von beiden Methoden für eine Residuenzahl kleiner als fünf beobachtet. Ausgehend von den GBSA-Berechnungen scheint aber die Additivität für Helices erfüllt zu sein, die aus mehr als 10 Residuen bestehen. Daher ist Vorsicht angebracht, wenn es darum geht, SASA-Parameter, die auf der Grundlage von Löslichkeit oder Verteilungskoeffizienten von kleine Molekülen abgeleitet wurden, bei großen Systemen anzuwenden. Umgekehrt dürfte es ebenfalls problematisch sein, Werte aus großen Systemen auf kleine Moleküle anzuwenden. Das Design von vereinfachten Modellen, in denen Helices als „residue beads" repräsentiert und Interaktionen additiv modelliert werden, stellt eine Herausforderung dar.

Untersuchung der Protonierungsgleichgewichte von Aminsäureseitenketten-Analoga

Unter Verwendung von Q-HOP MD Simulationen untersuchten wir die Protonierungsgleichgewichte von Essigsäure und 4-Methlyimidazol in wäßriger Lösung mit benachbarten Protonen-Akzeptorgruppen. Im Verlauf der Simulation von solvatierter Essigsäure wurden zwei verschiedene Arten von Protontransfers beobachtet. Ausgedehnte Phasen von häufigem Protonensprüngen zwischen Essigsäure und benachbarten Wassermolekülen können unterschieden werden von Phasen, in denen das Proton sich frei in der Simulationsbox bewegt, bis es wieder von einem Essigsäuremolekül eingefangen wird. Der pK_a wurde mit einem Wert berechnet, der etwa bei 3.0 liegt und basiert auf dem relativen Bestand von protonierten und deprotonierten Zuständen und dem Diffusionskoeffizienten von überschüssigen Protonen. Beide Werte stimmen gut mit experimentellen Messungen überein. Während der Untersuchung von 4-Methylimidazol in wäßriger Lösung mit benachbarten Protonakzeptorgruppen wurde eine qualitativ unterschiedliche Protonierung von 4-Methylimidazol im Vergleich zur Essigsäure beobachtet. Wegen des relativ hohen pK_a neigt $4MIH^+$ stark dazu, an einem Proton festzuhalten, sobald es gebunden ist. Eine nahegelegene titrierbare Gruppe mit niedrigerem pK_a-Wert hat nur wenig Chancen, das Proton von 4MIH zu übernehmen, die von äußeren Einflüssen bedingt sind. 4MI hat jedoch einen relativ kleinen Protoneneinfangradius, wodurch aus größeren Entfernungen nur schwerlich Protonen angezogen werden können. Protonierte Essigsäure kann sich das Proton leicht mit benachbarten titrierbaren Gruppen teilen, selbst wenn der Akzeptor nur einen geringen pK_a-Wert aufweist. Darüberhinaus hat AC^- einen weiten Protoneneinfangradius (ca. 5Å), der ihn zu einem perfekten „Protonenfänger" macht. Wasserstoffbrückenbindungen, an denen Aminosäureanaloga, Histidin und Asparaginsäure beteiligt sind, kommen häufig an Protontransfer-Pfaden vor. Wir sind der Meinung, dass die Ergebnisse aus den Studien zu den Verbindungen 4MI und ACH für den biologischen Protonentransfer von Bedeutung sind und zum Gegenstand künftiger Forschungsarbeit werden.

Chapter 1

Introduction

1.1 Molecular simulations — more than 50 years old but still young

1.1.1 The birth of molecular simulation

In 1954, Metropolis and his co-workers established the methodology of modern Monte Carlo (MC) simulation (*1*). The first computer simulation of a molecular liquid system was carried out using this very first MC technique at the Los Alamos National Laboratory in USA. In this model, the molecular systems were treated as hard spheres and disks. Therefore the results obtained from these simulations are highly idealized. However, this model was soon extended by adopting the Lennard-Jones non-bonded interaction potential (*2*), which made it possible to generate data that can be compared to experimental measurements. The MC simulation technique generates different configurations for a given system by making random changes in the position of every "particle" in the system, together with changes in its orientation and conformation, where appropriate. In a MC simulation the outcome of each trial movement of the "particle" depends only on its immediate predecessor, therefore there is no temporal relationship between successive MC configurations (*3*). As a result, time dependent properties, e.g. dynamics, of the system cannot be derived from a MC simulation. In order to obtain the dynamic properties of the studied system, a different simulation technique is needed. The Molecular dynamics (MD) simulation method generates successive configurations of the system by integrating Newton's laws of motion. The changes in the positions and velocities of every "particle" in the system are recorded in a trajectory. MD simulations calculate the "real" dynamics of the system, from which time averages of each property can be computed (*3*). The first MD simulation of a system in condensed phase was performed by Alder and Wainwright in 1957 (*4, 5*). In this model, a hard-sphere potential was used and the spheres moved at constant velocity in straight lines between collisions. Seven years later, a more realistic model of intermolecular interactions, i.e. a continuous potential, was developed by Rahman (*6, 7*). In this model, the force on a particle changed according to the change in the positions of the particle, or the changes of the position of any of the particles within the interaction range of this particle. Since the force changes continuously during the calculation, step-by-step techniques such as the *finite difference method* replaced the analytical solutions. In the 70's to 80's, many other methods and techniques were

developed for MD simulations making it a powerful tool in biophysics, material sciences as well as in energy chemistry.

1.1.2 Computer simulation of biomolecules

The first molecular dynamics simulation of a biological macromolecule was performed in 1977 by McCammon and Karplus on the bovine pancreatic trypsin inhibitor (BPTI) (*8*). *"The results were instrumental in replacing our view of proteins as relatively rigid structures"* (*9*). Even though the simulation seems very crude (vacuum environment, only 9.2 ps in length) by today's standards, it opened the gate for probing the dynamics of large biomolecules using computer simulation methods. To account for the importance of the presence of solvent in stabilizing the folded structure of biomolecules, an extension of this simulation was done by applying a simple spherical solvent model, which interacts with the protein only through the van der Waals interactions (*10*). This first MD simulation of nucleic acids was published in 1982. The simulation was also carried out in a vacuum environment due to the lack of a proper solvent model and the limited computational resources (*11*). Several solvent models have been developed in the 80's for a more reliable treatment of the solvation contributions. Some of the most successful ones are the simple point charge (SPC) model and its extension SPC/E model (*12, 13*), the transferable intermolecular potential with three particles (TIP3P) and its variant TIP4P (*14*). Based on these developments, the simulation of proteins or/and pieces of DNA with explicit descriptions of water molecules was soon made possible (*15-17*).

Molecular dynamics simulations can provide the ultimate detail concerning individual particle motions as a function of time. Moreover, acquiring the properties of a model system is often easier than experiments on the actual system. Therefore, in today's research, computer simulations are used to address many specific questions and details that are of interest for many biomolecular functions. In general, there are three types of applications of simulation methods in the macromolecular area (*9*), including 1) using simulations as a means of sampling configuration space; 2) using simulations to obtain a description of the system at equilibrium, including structural and motional properties and to calculate thermodynamic parameters like free energy changes, heat capacities, etc; 3) using simulations to study the dynamics of a system. These three types of applications require increasing demands on simulation methods (length and precision): all three types of application need adequate sampling of the configuration space. The latter two need also the condition that each point in the sampling is weighted by the appropriate Boltzmann factor. The third area even requires a correct representation of the development of the system over time. For the first two types, both molecular dynamics and Monte Carlo simulations can be used, while for the third area only MD can provide useful information.

1.2 Assumptions in molecular mechanics — force fields and its limit

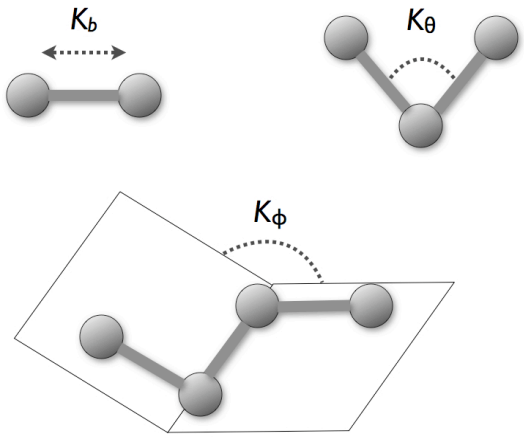
1.2.1 Molecular mechanics force fields

The foundations of molecular dynamics are given by the Born-Oppenheimer approximation that states that since the masses of the nuclei are much greater than the masses of the electrons, the electronic wave function thus depends only on the positions of the nuclei and not on their momenta. I.e. the movement of electrons and atoms can be treated separately, and the atoms can be well represented as classical point particles that follow a classical Newtonian dynamic. In classical molecular mechanics, the effect of the electrons is approximated as an effective potential function, the parameters of which are usually derived through fitting to more accurate methods or experimental properties. In MD simulations, the potential function is a combination of terms by which the particles in the simulation will interact. This is usually referred to as a force field. The most widely used molecular mechanical force fields, e.g. AMBER (*18*), CHARMM (*19*), GROMOS (*20*) and OPLSAA (*21*) incorporate a simple potential energy function U of the three dimensional coordinates q of the system, adjusting a large number of parameters to optimize the agreement with experimental data and with *ab initio* calculations on small molecules:

$$\begin{aligned} U(q) = & \sum_{bonds} \frac{1}{2} K_b (b - b_0)^2 + \sum_{bond\ \ angles} \frac{1}{2} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{torsional\ \ angles} \frac{1}{2} K_\phi [1 - \cos(n\phi + \delta)] \\ & + \sum_{atom} \sum_{pairs} \frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^6} \\ & + \sum_{atom} \sum_{pairs} \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q_i q_j}{r_{ij}} \end{aligned} \tag{1}$$

The first three terms here (from top to bottom) are bond stretching (1, 2 interaction), bond angle bending (1, 3 interaction) and bond rotation or torsion (1, 4 interaction, see Scheme 1). The last two terms describe the interactions between non-bonded pairs (in some case 1, 4 pairs are also included with separate parameters), which includes the dispersion attractions and exchange repulsion that are formulated as a Lennard-Jones function and the electrostatic interactions between atomic partial charges following Coulomb's law.

The potential function representing the non-bonded interactions in this example is a pair potential, in which the total potential energy of a system can be calculated from the sum of energy contributions between pairs of atoms. Since the non-bonded interactions are non-local and involve weak interactions between every pair of particles in the system, they are normally the bottleneck in the speed of MD simulations. If the system contains N particles, the computational cost is $O(N^2)$. Current available treatments of the non-bonded electrostatic interactions are numerical approximations such as cutoff, Reaction Field algorithms (*22, 23*), Particle Mesh Ewald summation (PME) (*24*), or the newer Particle-Particle Particle-Mesh (P3M) (*25*). These approximations can achieve a computational cost of $O(N \log N)$, which is a great improvement for large systems that are being studied now-a-days.



Scheme 1: Bonded interactions in molecular mechanics force fields

1.2.2 Molecular dynamics simulations

In molecular dynamics, successive configurations of the system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time (*3*). Newton's equations of motion can be written as follows:

$$\frac{dq_i(t)}{dt} = v_i(t) \tag{2}$$

$$\frac{dv_i(t)}{dt} = \frac{f_i(t)}{m_i} \tag{3}$$

$$f_i(t) = \frac{\partial}{\partial q_i} U(q_1, q_2, \ldots q_N) \tag{4}$$

Here f_i is the force on a particle with mass m_i and coordinate q_i, v_i is the velocity of particle i. Equations 2–3 are first-order differential equations. A second-order differential equation can be obtained by substituting eq. 2 into eq. 3 (*26*):

$$\frac{dq_i^2(t)}{dt} = \frac{f_i(t)}{m_i} \tag{5}$$

In order to perform a molecular dynamics simulation, the equations of motion must be solved for each particle. This falls into the problem of integration of a set of ordinary differential equations, for which a variety of algorithms exists. For high accuracy solutions of the equations of motion, it is advantageous to solve the sets of first-order equations for each particle. However, it is more efficient to solve the second-order differential equations for normal use due to the special form of Newton's equation (*26*). Methods for solving these equations are generally called *Verlet methods* (or *Verlet algorithm*) after L. Verlet, who belongs to the pioneers of applying integration algorithms to molecular simulations.

The *Verlet algorithm*: suppose at time t, the positions of the particles in the system are $\boldsymbol{R}(t)$, then the positions of the particles at time $t+\Delta t$ can be obtained from a Taylor expansion in terms of the time interval, Δt, and the positions and their derivatives at time t:

$$\boldsymbol{R}(t+\Delta t) = \boldsymbol{R}(t) + \Delta t\dot{\boldsymbol{R}}(t) + \frac{(\Delta t)^2}{2}\ddot{\boldsymbol{R}}(t) + O((\Delta t)^3). \tag{6}$$

In the same fashion, the positions at time $t-\Delta t$ can be derived as:

$$\boldsymbol{R}(t-\Delta t) = \boldsymbol{R}(t) - \Delta t\dot{\boldsymbol{R}}(t) + \frac{(\Delta t)^2}{2}\ddot{\boldsymbol{R}}(t) - O((\Delta t)^3). \tag{7}$$

Adding eq. 6 and eq. 7 and using eq.5 gives an expression for the positions of the particles at time $t+\Delta t$ as a function of the positions and forces at the earlier time step:

$$\begin{aligned}\boldsymbol{R}(t+\Delta t) &= 2\boldsymbol{R}(t) - \boldsymbol{R}(t-\Delta t) + (\Delta t)^2\ddot{\boldsymbol{R}}(t) + O((\Delta t)^4) \\ &\cong 2\boldsymbol{R}(t) - \boldsymbol{R}(t-\Delta t) + (\Delta t)^2\boldsymbol{M}^{-1}\boldsymbol{F}(t)\end{aligned} \tag{8}$$

Subtracting eq. 7 and eq. 6 gives an expression for the velocities of the particles $\boldsymbol{V}$ at time t:

$$\mathbf{V}(t) = \dot{\mathbf{R}}(t) \cong \frac{1}{2\Delta t}(\mathbf{R}(t+\Delta t) - \mathbf{R}(t-\Delta t)) \tag{9}$$

In principle, eqs. 8 and 9 are sufficient for the integration of the equation of motions. However in practice, due to the fact that the velocities are unknown at time t before deriving the positions at $t+\Delta t$, it is slightly inconvenient to directly use these formulas. There are several ways to avoid this problem. The methods used most often are the *velocity Verlet method*

$$\boldsymbol{R}(t+\Delta t) = \boldsymbol{R}(t) + \Delta t\boldsymbol{V}(t) + \frac{(\Delta t)^2}{2}\boldsymbol{M}^{-1}\boldsymbol{F}(t) \tag{10}$$

$$\boldsymbol{V}(t+\Delta t) = \boldsymbol{V}(t) + \frac{\Delta t}{2}\boldsymbol{M}^{-1}(\boldsymbol{F}(t) + \boldsymbol{F}(t+\Delta t)) \tag{11}$$

and the so-called *leapfrog algorithm*:

$$\boldsymbol{V}(t+\frac{\Delta t}{2}) = \boldsymbol{V}(t-\frac{\Delta t}{2}) + \Delta t\boldsymbol{M}^{-1}\boldsymbol{F}(t) \tag{12}$$

$$\boldsymbol{R}(t+\Delta t) = \boldsymbol{R}(t) + \Delta t\boldsymbol{V}(t+\frac{\Delta t}{2}). \tag{13}$$

Here the velocities at time t can be computed as:

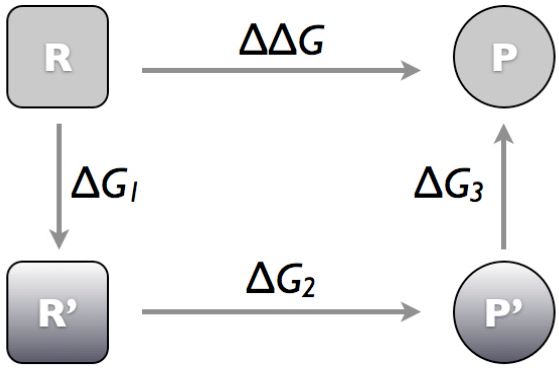
$$\boldsymbol{V}(t) = \frac{1}{2}(\boldsymbol{V}(t-\frac{\Delta t}{2}) + \boldsymbol{V}(t+\frac{\Delta t}{2})) \tag{14}$$

1.2.3 Successes and limitations of molecular mechanics force fields

In recent years, the general availability of molecular simulation packages has made molecular dynamics simulation a popular and powerful tool for the study of biological macromolecules in atomic detail. Compared to the first MD simulation of BPTI (about 500 atoms in size, less than 10 ps in length), current computer power allows simulations at 100 ns to microsecond scale of systems of 10^5 to 10^6 atoms in size. People are now studying much more complicated biological processes like water permeation across aquaporin (*27*), dynamics of a complete virus (*28*) as well as proton / sodium ion exchange in Na^+/H^+ antiporters (*29*).

Molecular dynamics simulation has been successfully used in: folding of short peptides (*30*), long range motions of functional domains in proteins or protein complexes (*31*), protein-ligand and protein-protein binding and interactions (*32, 33*), as well as calculation and interpretation of the relaxation time and NOE in nuclear magnetic resonance experiments. Despite the above-mentioned success of molecular dynamics simulation combined with molecular mechanics force fields, there are still some important areas where MD simulation cannot provide enough information. One example is chemical reactions, especially those reactions involving complicated environmental contributions such as enzymatic reactions or long-range proton transfer in biological / chemical systems. In these cases, the development of mixed quantum mechanical / molecular mechanic (QM/MM) methods has shone some light in the study of enzymatic reaction (*34*), where the reactions mostly take place among localized segments of a large molecule. However, this approach is still problematic in processes involving a large part of the system such as long-range proton transfer. In this thesis, we attempt to give some possible solutions to such problems within the MD regime.

1.3 Free energy calculation



Scheme 2: Thermodynamic cycle

Most of the important chemical quantities like binding affinity, association and dissociation constants, solubilities, as well as chemical potentials are directly related

to the free energy (or free energy change) of a molecular system. Therefore, the calculation of free energy (both Gibbs free energy and Helmholtz free energy) is of high importance for computational chemistry and computational biophysics. The statistical equilibrium averages can be obtained from an MD simulation for any desired property of the molecular system for which a value can be computed at each point of the trajectory. For example, the potential and kinetic energy of relevant parts of the system, structural properties and fluctuations, electric fields, diffusion constants etc, can all be derived from the simple average over a MD trajectory. However, the entropy and the free energy cannot be derived from such statistical averages. The dependence on the extent of accessible phase space makes it generally impossible to compute the absolute free energy of a molecular system. In the 80's to 90's, statistical mechanical procedures have been developed by means of which *relative* free energy differences may be obtained (*35*) (see Scheme 2). The thermodynamic cycle integration technique allows the calculation of such *relative* free energy differences. One of the most successful and well-established methods is called Multiconfiguration thermodynamic integration (MCTI) (*36*).

In MCTI, the free energy difference between two states A and B of a system is determined from a MD simulation, in which the potential energy function U is slowly changed such that the system slowly converts from state A into state B (*36*). The potential energy function U is expressed as a function of some control variable λ that determines the state of the system. As a consequence of the Hamiltonian H being a function of this control variable λ, the partition function Δ for the system is a function of λ as well. For an isothermal isobaric ensemble the partition function is

$$\Delta(\lambda) = \frac{1}{h^{3N}N!}\iiint \exp\left\{-\frac{H(\lambda)+pV}{k_BT}\right\}dVdp^Ndq^N. \qquad (15)$$

Here N is the number of particles, h is the Planck's constant, p is the pressure, V is the volume, T is the absolute temperature, H is the Hamiltonian, k_B is the Boltzmann's constant, and p^N and q^N are the momenta and positions of the N particles.

For an isothermal isobaric ensemble, the Gibbs free energy G is also a function of λ,

$$G(\lambda) = -k_BT\ln\Delta(\lambda) \qquad (16)$$

and the derivative with respect to λ is

$$\frac{\partial G(\lambda)}{\partial\lambda} = \left\langle\frac{\partial H(\lambda)}{\partial\lambda}\right\rangle_\lambda \qquad (17)$$

This is an ensemble average of $\partial H(\lambda)/\partial\lambda$ for a system with Hamiltonian $H(\lambda)$, which can be obtained directly from MD simulations. The free energy difference between two states of a system ($\lambda = 0$ and $\lambda = 1$) described by their Hamiltonians $H(\lambda = 0)$ and $H(\lambda = 1)$ can be then obtained by:

$$\Delta G = \int_0^1\left[\frac{\partial G(\lambda)}{\partial\lambda}\right]_\lambda d\lambda = \int_0^1\left\langle\frac{\partial H(\lambda)}{\partial\lambda}\right\rangle_\lambda d\lambda \qquad (18)$$

In practice, the integral has to be evaluated as a sum of ensemble averages due to the fact that the MD simulations are performed with discrete steps (windows) (*36*),

$$\Delta G = \sum_{i=1}^{n} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} \Delta\lambda_i \tag{19}$$

where n is the number of different values of λ, and $\Delta\lambda_i$ is the difference between successive values of λ. The estimate for the statistical error of the entire MCTI can be found by:

$$E(\Delta G) = \sqrt{\sum_i E_i^2 \Delta\lambda_i}\ , \tag{20}$$

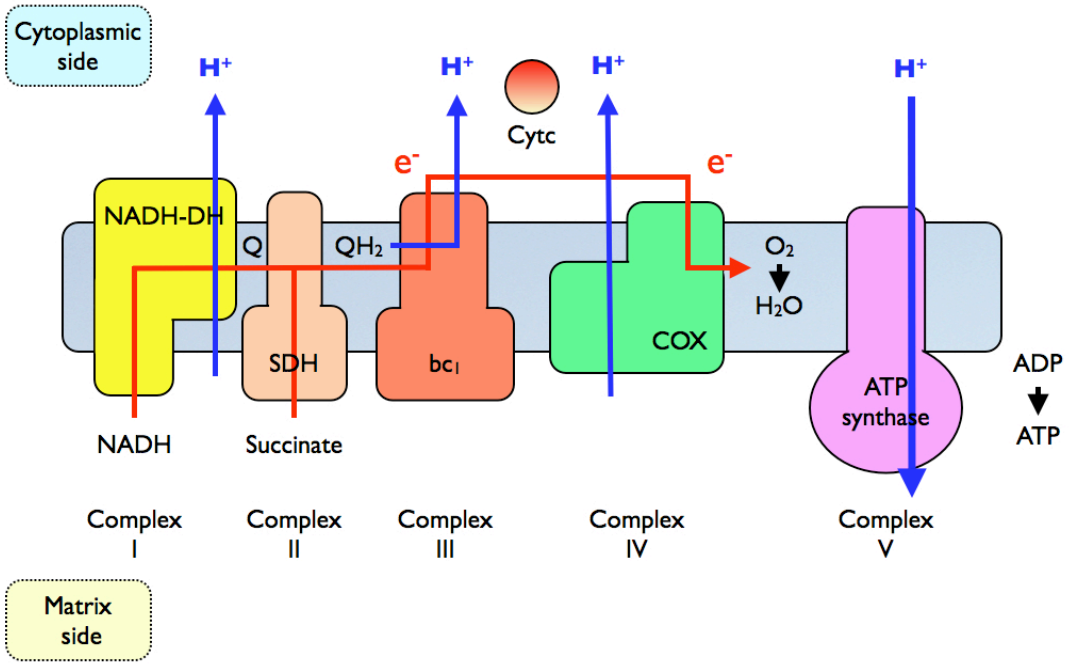
where $E_i = E\left(\left\langle \partial H(\lambda)/\partial\lambda \right\rangle_i \right)$ is the statistical error at each value of λ_i (*36*).

1.4 Proton transfer as an example — an attempted improvement of the state-of-the-art MD

1.4.1 Proton transfer process and its importance in chemistry and biology

Proton transfer is essential in many biological and chemical processes. For example, cellular proton pumps such as cytochrome *c* oxidase (*37, 38*) and the cytochrome bc_1 complex (*39, 40*) generate proton gradients across biological membranes, which are then used by other biological processes such as for the synthesis of ATP (see Scheme 3). Proton transfer reactions are also crucial in other areas, for example, for membrane permeation in hydrogen fuel cells or in polymers (*41*). Furthermore, the protein structure itself is often strongly dependent on the predominant protonation states of the titratable side chain groups as well (*42, 43*).

In spite of their enormous importance, many aspects of proton transfer (PT) reactions in biomolecules remain poorly understood. Experimental techniques, in particular, are facing fundamental and / or technical difficulties with respect to the direct observation of PT reactions. For example, X-ray crystallography is widely used in determining the three-dimensional structures of biological macromolecules. Nevertheless, hydrogen atoms cannot be detected in most structures except for a few structures at ultrahigh resolution. Although NMR experiments can detect protons directly, the time resolution of NMR is not short enough to resolve proton transfer processes which occur on time scales as short as tens of femtosecond. Similarly, neutron diffraction experiments can only provide time-averaged proton positions. Mass spectroscopy experiments need to be performed under vacuum conditions and often face the problem of proper peak assignment. Apparently, the only direct experimental observation of PT reactions is from Fourier Transform Infrared (FTIR) spectroscopy that is able to identify proton transfer paths implicitly when combined with site-directed mutagenesis (*44*). Therefore, it is highly desirable to complement the existing experimental techniques by computational methods.



Scheme 3: Proton transfer (blue arrows) in the respiration chain of mitochondria.

1.4.2 Current theories and models of proton transfer

In the past decades, various computational methods have been developed to calculate the pK_a values of amino acid side chains as well as to perform constant pH simulations of proteins (*45-51*). Hünenberger and co-workers proposed a model using "fractional charges" that allows continuous changes between different protonation states (*45*). Such kinds of fractional models (*45, 46*), however, have also been criticized because of their nonphysical intermediate protonation state (*47, 48*). The work of Brooks and co-workers addressed this problem using a set of continuous titration coordinates that describes transitions between fully protonated or deprotonated states (*49*). Besides the continuous models, several discrete models were proposed by combining Monte Carlo sampling for selecting the protonation states with Poisson-Boltzmann methods (*50, 51*) or thermodynamic integration (*52*) for calculating protonation energies. Recently, Mongan et al. introduced an efficient model that uses the generalized Born (GB) implicit solvation model for the protonation state transition energies and dynamics (*48*). The methods mentioned above allow computing the pK_a of amino acid side chains in a relative efficient manner. However, they do not model explicit proton exchange reactions between the titratable sites and the surrounding aqueous solution or the exchange between different titratable sites. Therefore, these methods may not be suitable to identify proton transfer pathways or to characterize the mechanisms of PT reactions.

This is the area where dynamic simulations of proton transfer come into play. Tuckerman; Marx and co-workers studied the shared proton in hydrogen bonds (*53*) and a hydrated excess proton in water (*54*) using the Car-Parrinello molecular dynamics (CPMD) method (*55, 56*). Lobaugh and Voth investigated proton transport in water by simulating an excess proton in a box of water molecules (*57*) within the

centroid molecular dynamics (CMD) (*58*) framework. A similar system was then studied using a multistate empirical valence bond (MS-EVB) model for proton transfer (*59-61*). A recent study presented the dynamic simulation of pK_a values for amino acid side analogues (*62*) using the MS-EVB model and the umbrella sampling technique (*63, 64*). There, different parts of the system phase space were sampled by fixing the distance between the donor and the acceptor (distance between center of excess charge) at different values. The deviation between their computed value and the experimental pK_a was 1-2 pK_a units. Voth and co-workers also studied aqueous proton solvation and transport using the CPMD method (*65*). Besides such model systems, several applications showed the importance and success of studying proton transfer in protein systems by theoretical approaches, for instances, the proton shuttle in green fluorescent protein (*66*), the proton transfer in bacteriorhodopsin (*67*), the proton transfer in Gramicidin A (*68, 69*), the proton transfer along a water chain in the D-pathway of COX (*70*) and the proton translocation in Carbonic Anhydrase (*71*).

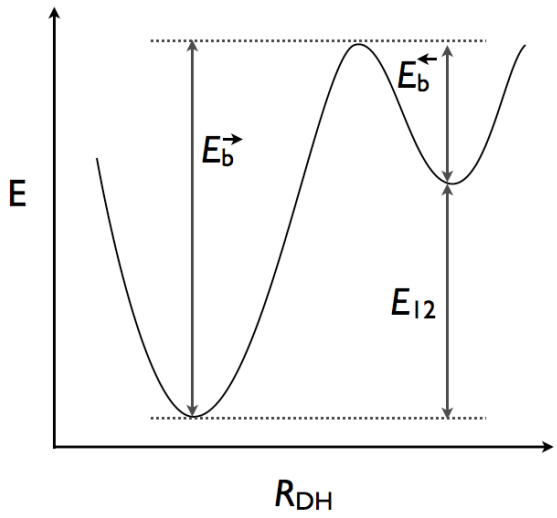
In biological systems, especially in membrane proton pumps, the proton transfer pathways may extend over several nanometers involving many titratable amino acid as well as water chains (*72*). The mechanisms of such long transfer processes are, therefore, rather complicated (*38*) and the principles behind these long distance proton-transport processes are best revealed by transferring knowledge obtained on well-understood model systems or subprocesses of the more complex systems. The studies mentioned above provided quantum or semiempirical descriptions of short-range proton transfer process or described the diffusion of excess protons in bulk water. However, no such study has so far addressed the long term diffusion of protons involving both amino acids and solvent environment. Although several studies have successfully combined quantum mechanical and molecular mechanics force field approaches (QM/MM) with path-sampling techniques (*67, 73, 74*) to study proton transfer in biological macromolecules, there is certainly a great need for semi-quantitative approaches that can efficiently explore proton transfer paths in proteins consisting of many subsequent transfer events. Once these transfer paths are identified, more accurate and well-established methods can be applied to compute PMFs of individual reaction steps.

One such model of intermediate accuracy, the Q-HOP MD method, was introduced earlier to study proton transport in biomolecular systems (*75-78*). In the Q-HOP scheme, the dynamics of a classical simulation system is propagated by conventional Newtonian molecular dynamics and stochastic proton transfer events allow for dynamic protonation changes of the titratable groups in the system. The corresponding proton transfer events are abstracted as quasi-instantaneous hopping from a donor group to an acceptor group. In this way, the total number of protons is conserved. The proton transfer likelihood per time step, termed the proton transfer probability, is calculated on the fly for each proton donor and acceptor pair using a parameterized functional form during the MD simulation. Depending on whether the proton transfer takes place or not (by comparing the proton transfer probability with a random number), the topology of the system will be modified or be kept unchanged before the next step of MD simulation. The transfer probabilities depend on the actual donor-acceptor distance, termed RDA, and the energy difference between the two minima at donor and acceptor, termed E_{12} in the momentary configuration. In contrast to MS-EVB and ab initio based molecular dynamics simulation methods, the Q-HOP method does not include an explicit treatment of a delocalized proton. These details

are believed to be of less importance for identification of transfer pathways. In protonation equilibria as studied in this work, a shared proton between donor and acceptor would be reflected by frequent exchanges between both groups. The Q-HOP MD method has been successfully applied to study the proton shuttle in green fluorescent protein (GFP) (*66*), to understand the mechanism of proton blockage in Aquaporin (*79*), and to study the explicit protonation equilibrium of solvated amino acid analogues on a time scale of tens of nanoseconds at low p*H* conditions (*80*).

1.4.3 The Q-HOP method of dynamic simulation of proton transfer

In the Q-HOP method, the proton hopping probability p is calculated from the energy difference E_{12} (see Scheme 4) between the pair of protonated donor/deprotonated acceptor and the pair of deprotonated donor/protonated acceptor, and from the distance between donor and acceptor atoms, R_{DA} (*78*), see equations (24)-(26) below. Depending on the values of E_{12} and R_{DA}, two different approaches are used to compute the hopping probability (transfer rate) (*76*).



Scheme 4: Illustration of some important quantities in the Q-HOP model.

For large E_{12} and large R_{DA}, a modified transition state theory is used accounting for the zero-point energy and the tunneling effect:

$$p = \kappa(T, E_M)\frac{k_B T}{h}\exp\left(-\frac{E_b - h\omega/2}{k_B T}\right)\Delta t \qquad (21)$$

where $\kappa(T, E_M)$ is the enhancement of the classical transfer rate due to tunneling $\kappa = k_{QM}/k_{classical}$ as function of temperature T and E_M (*76*). $E_M = E_{\max} - \max(E_{\min,1} - E_{\min,2})$ is the difference between the energy maximum $E_{\max}$ and the larger one of the two energy minima $E_{\min,1}$ and $E_{\min,2}$ along the two-well potential of a typical transition reaction. $h\omega/2$ is the zero-point energy obtained by considering the bonds that contain a transferring proton as a quantum-mechanical

harmonic oscillator with frequency ω at the educt well minimum (*78*). E_b is the energy barrier along the two well potential and is calculated in Q-HOP as a function of E_{12} and R_{DA}:

$$E_b(E_{12}, R_{DA}) = S(R_{DA}) + T(R_{DA})E_{12} + V(R_{DA})E_{12}^2 \quad (22)$$

where *S*, *T* and *V* have a simple functional dependence on R_{DA} (*42*).

In practice, this high-barrier regime is only a limiting case. In nanosecond time-scale simulations, the probabilities computed from eq (21) are too low for PT events to occur. For barriers involving small E_{12} and small R_{DA}, the transfer rate is estimated by following the propagation of a one-dimensional wave package as a solution of the time-dependent Schrödinger equation and computing the fractional population that crosses the barrier:

$$p = 0.5\tanh(-K(R_{DA}))E_{12} + M(R_{DA}) + 0.5 \quad (23)$$

where *K* and *M* are also functions of R_{DA} (*76*).

In the Q-HOP method, E_{12} is a sum of two contributions:

$$E_{12} = E_{12}^0 + E_{12}^{env} \quad (24)$$

E_{12}^0 is the energy difference between a donor-acceptor pair in vacuum. It is obtained from the following empirical relationship (*78*),

$$E_{12}^0 = \alpha + \beta \cdot R_{DA} + \gamma \cdot R_{DA}^2 \quad (25)$$

where α, β and γ are fitted parameters compiled in a recent data set involving MP2/6-31++G** calculations of all titratable amino acids (*81*). The environmental contribution E_{12}^{env} is calculated from the coulombic interactions between the two pairs and the environment:

$$E_{12}^{env} = E_{DH-A,env}^{Coul} + E_{D-AH,env}^{Coul} \quad (26)$$

Here $E_{DH-A,env}^{Coul}$ and $E_{D-AH,env}^{Coul}$ are defined as:

$$E^{Coul} = \sum_{i}^{donor\_acceptor\_atoms} \sum_{j}^{remaining\_system} \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \quad (27)$$

q_i and q_j are the respective atomic partial charges, r_{ij} is the atomic distance between atoms *i* and *j*, and ε_0 is the permittivity of vacuum. The atomic partial charges are obtained from an optimization procedure to reproduce E_{12} energies from QM/MM calculations. These charges are not used in the propagation of the trajectory (normal MD part), but only serve to compute the environmental contribution from equation (26).

1.5 Goals of this thesis

1.5.1 Studying the interaction between proline-rich peptides and their adapter domain (Chapter 2 – 5)

Our group was involved in a collaboration with the experimental group of Dr. Christian Freund at Freie Universität Berlin, who is using NMR spectroscopy and phage display in combination with SPOT analysis to study the peptide-protein interactions. In the context of a project funded by the Volkswagen foundation, we were interested in the modes of interactions between the glycine-tyrosine-phenylalanine (GYF) adaptor domains and proline-rich peptides. The recognition of proline-rich sequences plays an important role for the assembly of multi-protein complexes during the course of eukaryotic signal transduction and is mediated by a set of protein folds that share characteristic features (*82*). The GYF domain is known as a member of the super-family of recognition domains for proline-rich sequences. The role of the simulation partner in this project was to model the solvent structure of wild-type and mutant polyproline peptides and present structural models of different complexes that can explain the binding motifs obtained by the experimental partners at atomic level. The original plan also included the design of a virtual screening method for large scale screening of interesting peptide motifs that interact with the GYF domain.

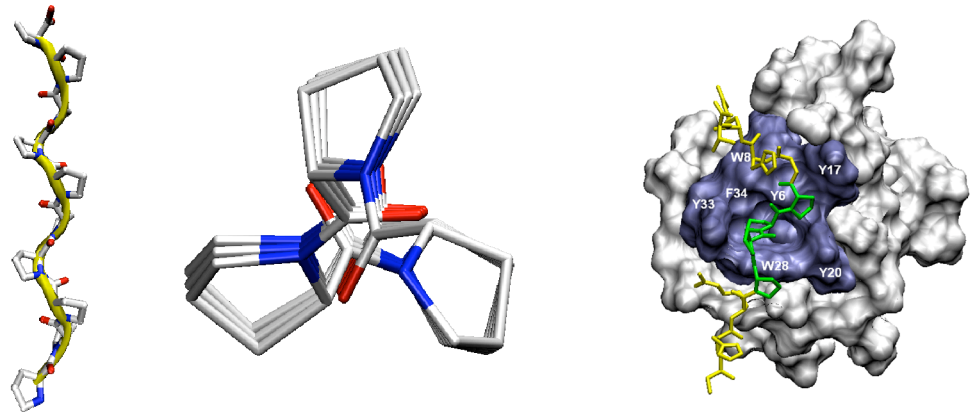


Figure 1: The PPII helices and its view along the helical axis (left and middle); NMR structure of the GYF domain with wild-type proline-rich peptide (right).

By molecular modeling and MD simulations, we studied the solvent conformation of the wild-type and mutant polyproline peptides that bind to the GYF domain. We found that the peptides formed PPII helix (see Figure 1) conformations even in the absence of the GYF domain. These results agree well with recent experimental and theoretical studies on polypeptides with or without prolines and indicate that the formation of a PPII helix of the peptide is not induced by the binding processes alone. Based on our previous knowledge from NMR experimental studies of the GYF domain-ligand interaction and the simulations of the wild-type and mutated complexes, we modeled the general binding mode of polyproline peptides to the GYF domain. The hydrophobic interactions between the peptide residues Pro6 and Pro7, and the binding pocket as well as the electrostatic attractions between the

peptide residues Arg3 and Arg10, and the domain residues Glu31 and Glu9 play crucial roles in the binding. Peptide docking and subsequent MD simulations of the G8X mutants identified an alternative binding mode, where a shift in register for the interacting prolines was observed. These results agree qualitatively well with NMR chemical shift mapping experiments and indicate that dynamic processes are important for proline-rich sequence recognition. Possibly, such gliding motions along long proline-rich sequences decrease the entropic penalty of binding while still keeping a certain degree of specificity.

As a side project with Dr. Freund's group, we were also involved in the study of linear peptides binding to Cyclophilin A, which is a peptidyl-prolyl *cis*/*trans*-isomerase that is involved in multiple signaling events of eukaryotic cells (see chapter 3). Using peptide docking and MD simulation methods, we studied the binding of linear peptide motifs to Cyclophilin A. From substitution analysis (phage display), the linear sequence recognition code for CypA and the consensus motif FGPXLp were identified (done by our experimental collaborators). The modeled complex structure (docking + MD) agrees very well with the results from phage display experiments and gives an explanation of the specific binding motif from structural and interaction points of view.

Chapters 2 and 4 presented our work on individual complexes of proline-rich peptides binding to the GYF domain at atomic detail using MD simulations. However, for the fast screening of many peptide sequences, a more efficient method is needed. Docking trials by Oliver Müller and Cosima Graf (ex-bachelor students of Dr. Helms' group) with the docking tool FlexX (*83*) partly failed because FlexX does not include a solvation term in its energy function. Therefore we investigated whether peptide solvation can be treated in an approximate fashion. We calibrated this residue-scale model against atomistic free energy simulations of peptide helices in a solvent box, where during the simulation the peptide interactions are switched off. Chapter 5 presents calculations for the solvation free energies of α-helical peptides of various lengths by the MCTI method. In this study, non-additivity of the solvation free energy is found for peptides shorter than 5 residues. On the other hand, additivity appears fulfilled for longer helices. Thus, it is important to consider molecular details of backbone hydration, which is normally ignored in solvent models at residue-scale. The design of simplified models, where helices are composed of residue-beads and interactions are modeled additively, appears challenging.

1.5.2 Studying protonation equilibria of amino acid side-chain analogs (chapters 6 – 7)

The Helms group has a long-standing interest in studying proton migration in biomolecules. Previous Ph.D students of the Helms group (at MPI of Biophysics, Frankfurt) Marcus Lill, Tomaso Frigato and Elena Herzog developed a new method, namely Q-HOP MD simulation, that is able to study the proton transfer in large biological / chemical systems. When I entered into this project, the implementation of Q-HOP into the parallel package NWChem was completed and simulations over long timescales were suddenly possible. We are continuing to expand the applicability of the Q-HOP method to biomolecules to study the proton transfer processes in cytochrome *c* oxidase and fumarate reductase (together with Elena Herzog and Dr.

Roy Lancaster). As a control of the methodology, we attempted first to compute the pK_a equilibria of model substances from unbiased simulations.

Using the Q-HOP MD simulation technique, we studied the protonation equilibria of acetic acid (AC^-/ACH) and 4-methylimidazole ($4MIH^+$/4MI) in aqueous solution and with nearby proton accepting groups. In the simulation of solvated acetic acid (see Figure 2 and chapter 6), two different regimes of proton transfer were observed. Extended phases of frequent proton swapping between acetic acid and nearby water were separated by phases where the proton freely diffuses in the simulation box until it is captured again by acetic acid. The pK_a of acetic acid was calculated around 3.0 based on the relative population of protonated and deprotonated states and the diffusion coefficient of excess proton was computed from the average mean squared displacement in the simulation. Both calculated values agree well with the experimental measurements.

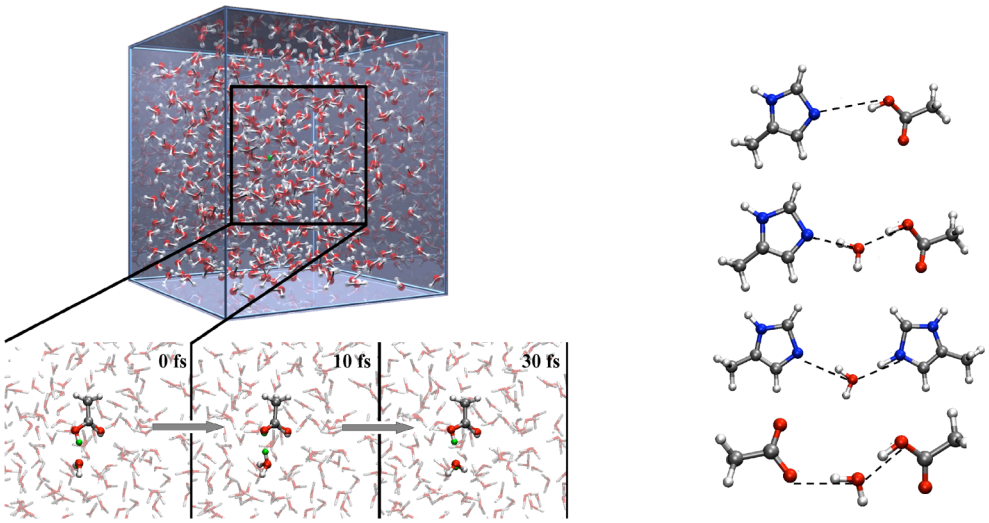


Figure 2: Schematic representations of the systems studied using the Q-HOP MD simulations. Left: solvated acetic acid. Right: 4-methylimidazole in aqueous solution and with nearby proton accepting groups.

In the studied of 4-methylimidazole in aqueous solution and with nearby proton accepting groups (see Figure 2 and chapter 7), qualitatively different protonation behavior of 4-methylimidazole compared to that of acetic acid was found: Due to its relatively high pK_a $4MIH^+$ has a high tendency to keep a proton once it is bound. A close titratable group with lower pK_a only has few chances to snatch the proton from $4MIH^+$ that are driven by environmental fluctuations. On the other hand, 4MI has a relatively small proton capture radius, making it very hard to attract protons from long distances. Protonated acetic acid can easily share the proton with close titratable groups even if the acceptor group has a low pK_a. Moreover, AC^- has a large proton capture radius (about 5 Å), making it a perfect proton "capturer". Hydrogen bond chains involving the amino acid analogs, histidine and aspartic acid, are frequently found along proton transfer pathways in biomolecules. We suggest that the findings of this study on the model compounds 4MI and ACH are of relevance to biological proton transfer and this will be addressed in future work.

References

1. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of State Calculations by Fast Computing Machines, *J. Phys. Chem. 21*, 1087-1092.
2. Wood, W. W., and Parker, F. R. (1957) Monte Carlo Equation of State of Molecules Interacting with the Lennard-Jones Potential. I. A Supercritical Isotherm at about Twice the Critical Temperature, *J. Phys. Chem. 27*, 720-733.
3. Leach, A. R. (2001) *Molecular modelling, principles and applications*, Pearson Education.
4. Alder, B. J., and Wainwright, T. E. (1957) Phase Transition for a Hard Sphere System, *J. Phys. Chem. 27*, 1208-1209.
5. Alder, B. J., and Wainwright, T. E. (1959) Studies in Molecular Dynamics. I. General Method, *J. Phys. Chem. 31*, 459-466.
6. Rahman, A., and Stillinger, F. H. (1971) Molecular Dynamics Study of Liquid Water, *J. Phys. Chem. 55*, 3336-3359.
7. Rahman, A. (1964) Correlations in the Motion of Atoms in Liquid Argon, *Phys. Rev. 136*, 405.
8. McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of folded proteins, *Nature 267*, 585-590.
9. Karplus, M., and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol. 9*, 646-652.
10. Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation, *Adv. Prot. Chem. 14*, 1-63.
11. Levitt, M. (1982) Computer-simulation of DNA double-helix dynamics, *Cold Spring Harbor Symp. Quant. Biol. 176*, 251-262.
12. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981) in *Intermolecular forces* (Pullman, B., Ed.) pp 331-342, Reidel, Dordrecht.
13. Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The Missing Term in Effective Pair Potentials, *J. Phys. Chem. 91*, 6269-6271.
14. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys. 79*, 926-935.
15. van Gunsteren, W. F., Berendsen, H. J. C., Hermans, J., Hol, W. G. J., and Postma, J. P. M. (1983) Computer Simulation of the Dynamics of Hydrated Protein Crystals and Its Comparison with X-Ray Data, *Proc. Natl. Acad. Sci. USA 80*, 4315-4319.
16. van Gunsteren, W. F., and Berendsen, H. J. C. (1984) Computer simulation as a tool for tracing the conformational differences between proteins in solution and in the crystalline state, *J. Mol. Biol. 176*, 559-564.
17. Seibel, G. L., Singh, U. C., and Kollman, P. A. (1985) A Molecular Dynamics Simulation of Double-Helical B-DNA Including Counterions and Water, *Proc. Natl. Acad. Sci. USA 82*, 6537-6540.
18. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995), *J. Am. Chem. Soc. 118*, 2309-2309.
19. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins, *J. Phys. Chem. B 102*, 3586-3616.
20. Lukas D. Schuler, X. D., Wilfred F. van Gunsteren,. (2001) An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase, *J. Comput. Chem. 22*, 1205-1218.
21. Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids, *J. Am. Chem. Soc. 118*, 11225-11236.
22. Tironi, I. G., Sperb, R., Smith, P. E., and Gunsteren, W. F. v. (1995) A generalized reaction field method for molecular dynamics simulations, *J. Phys. Chem. 102*, 5451-5459.

23. Barker, J. A., and Watts, R. O. (1973) Monte Carlo studies of the dielectric properties of water-like models, *Mol. Phys. 26*, 789 - 792.
24. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A Smooth Particle Mesh Ewald Method, *J. Chem. Phys. 103*, 8577-8593.
25. Hockney, R., and Eastwood, J. (1981) *Computer simulation using particles*, McGraw-Hill, New York.
26. Field, M. J. (1999) *A practical introduction to the simulation of molecular systems*, Cambridge University Press, Cambridge.
27. de Groot, B. L., and Grubmuller, H. (2001) Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF, *Science 294*, 2353-2357.
28. Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A., and Schulten, K. (2006) Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus, *Structure 14*, 437-449.
29. Arkin, I. T., Xu, H., Jensen, M. O., Arbely, E., Bennett, E. R., Bowers, K. J., Chow, E., Dror, R. O., Eastwood, M. P., Flitman-Tene, R., Gregersen, B. A., Klepeis, J. L., Kolossvary, I., Shan, Y., and Shaw, D. E. (2007) Mechanism of Na+/H+ Antiporting, *Science 317*, 799-803.
30. Zhou, Y., and Karplus, M. (1999) Interpreting the folding kinetics of helical proteins, *Nature 401*, 400-403.
31. Bockmann, R. A., and Grubmuller, H. (2002) Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase, *Nat. Struct. Biol. 9*, 198-202.
32. Gu, W., Kofler, M., Antes, I., Freund, C., and Helms, V. (2005) Alternative Binding Modes of Proline-Rich Peptides Binding to the GYF Domain, *Biochemistry 44*, 6404-6415.
33. Eyrisch, S., and Helms, V. (2007) Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction., *J. Med. Chem. 50*, 3457-3464.
34. Gao, J., and Truhlar, D. G. (2002) Quantum mechanical methods for enzyme kinetics, *Annu. Rev. Phys. Chem. 53*, 467-505.
35. van Gunsteren, W. F., and Berendsen, H. J. C. (1990) Computer simulation of molecular dynamics: methodology, applications, and perspective in chemistry, *Angew. Chem. Int. Ed. 29*, 992-1023.
36. Straatsma, T. P., and McCammon, J. A. (1991) Multiconfiguration thermodynamic integration, *J. Phys. Chem. 95*, 1175-1188.
37. Iwata, S., Ostermeier, C., Ludwig, B., and Michel, H. (1995) Structure at 2.8-Angstrom Resolution of Cytochrome-C-Oxidase from Paracoccus-Denitrificans, *Nature 376*, 660-669.
38. Michel, H. (1999) Cytochrome c oxidase: Catalytic cycle and mechanisms of proton pumping-A discussion, *Biochemistry 38*, 15129-15140.
39. Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S., and Jap, B. K. (1998) Complete Structure of the 11-Subunit Bovine Mitochondrial Cytochrome bc1 Complex, *Science 281*, 64-71.
40. Xia, D., Yu, C.-A., Kim, H., Xia, J.-Z., Kachurin, A. M., Zhang, L., Yu, L., and Deisenhofer, J. (1997) Crystal Structure of the Cytochrome bc1 Complex from Bovine Heart Mitochondria, *Science 277*, 60-66.
41. Jang, S. S., Lin, S. T., Cagin, T., Molinero, V., and Goddard, W. A. (2005) Nanophase segregation and water dynamics in the dendrion diblock copolymer formed from the Frechet polyaryl ethereal dendrimer and linear PTFE, *J. Phys. Chem. B 109*, 10154-10167.
42. Dlugosz, M., and Antosiewicz, J. M. (2005) The impact of protonation equilibria on protein structure, *J. Phys.: Condens. Matter 17*, S1607-S1616.
43. Alonso, D. O. V., DeArmond, S. J., Cohen, F. E., and Daggett, V. (2001) Mapping the early steps in the pH-induced conformational conversion of the prion protein, *Proc. Natl. Acad. Sci. USA 98*, 2985-2989.
44. Lecoutre, J., Tittor, J., Oesterhelt, D., and Gerwert, K. (1995) Experimental-Evidence for Hydrogen-Bonded Network Proton-Transfer in Bacteriorhodopsin Shown by Fourier-Transform Infrared-Spectroscopy Using Azide as Catalyst, *Proc. Natl. Acad. Sci. USA 92*, 4962-4966.
45. Borjesson, U., and Hunenberger, P. H. (2001) Explicit-solvent molecular dynamics simulation at constant pH: Methodology and application to small amines, *J. Chem. Phys. 114*, 9706-9719.
46. Baptista, A. M., Martel, P. J., and Petersen, S. B. (1997) Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration, *Proteins: Struct. Funct. Genet. 27*, 523-544.

47. Onufriev, A., Case, D. A., and Ullmann, G. M. (2001) A novel view of pH titration in biomolecules, *Biochemistry 40*, 3413-3419.
48. Mongan, J., Case, D. A., and McCammon, J. A. (2004) Constant pH molecular dynamics in generalized born implicit solvent, *J. Comput. Chem. 25*, 2038-2048.
49. Lee, M. S., Salsbury, F. R., and Brooks, C. L. (2004) Constant-pH molecular dynamics using continuous titration coordinates, *Proteins: Struct. Funct. Bioinfo. 56*, 738-752.
50. Walczak, A. M., and Antosiewicz, J. M. (2002) Langevin dynamics of proteins at constant pH, *Phys. Rev. E 66*, 051911.
51. Baptista, A. M., Teixeira, V. H., and Soares, C. M. (2002) Constant-pH molecular dynamics using stochastic titration, *J. Chem. Phys. 117*, 4184-4200.
52. Burgi, R., Kollman, P. A., and van Gunsteren, W. F. (2002) Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation, *Proteins: Struct. Funct. Genet. 47*, 469-480.
53. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1997) On the quantum nature of the shared proton in hydrogen bonds, *Science 275*, 817-820.
54. Marx, D., Tuckerman, M. E., Hutter, J., and Parrinello, M. (1999) The nature of the hydrated excess proton in water, *Nature 397*, 601-604.
55. Marx, D., and Parrinello, M. (1996) Ab initio path integral molecular dynamics: Basic ideas, *J. Chem. Phys. 104*, 4077-4082.
56. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1996) Efficient and general algorithms for path integral Car-Parrinello molecular dynamics, *J. Chem. Phys. 104*, 5579-5588.
57. Lobaugh, J., and Voth, G. A. (1996) The quantum dynamics of an excess proton in water, *J. Chem. Phys. 104*, 2056-2069.
58. Cao, J. S., and Voth, G. A. (1994) The Formulation of Quantum-Statistical Mechanics Based on the Feynman Path Centroid Density .1. Equilibrium Properties, *J. Chem. Phys. 100*, 5093-5105.
59. Schmitt, U. W., and Voth, G. A. (1998) Multistate empirical valence bond model for proton transport in water, *J. Phys. Chem. B 102*, 5547-5551.
60. Schmitt, U. W., and Voth, G. A. (1999) The computer simulation of proton transport in water, *J. Chem. Phys. 111*, 9361-9381.
61. Day, T. J. F., Soudackov, A. V., Cuma, M., Schmitt, U. W., and Voth, G. A. (2002) A second generation multistate empirical valence bond model for proton transport in aqueous systems, *J. Chem. Phys. 117*, 5839-5849.
62. Maupin, C. M., Wong, K. F., Soudackov, A. V., Kim, S., and Voth, G. A. (2006) A multistate empirical valence bond description of protonatable amino acids, *J Phys. Chem. A 110*, 631-639.
63. Patey, G. N., and Valleau, J. P. (1975) Monte-Carlo Method for Obtaining Interionic Potential of Mean Force in Ionic Solution, *J. Chem. Phys. 63*, 2334-2339.
64. Pangali, C., Rao, M., and Berne, B. J. (1979) Monte-Carlo Simulation of the Hydrophobic Interaction, *J. Chem. Phys. 71*, 2975-2981.
65. Izvekov, S., and Voth, G. A. (2005) Ab initio molecular-dynamics simulation of aqueous proton solvation and transport revisited, *J. Chem. Phys. 123*.
66. Lill, M. A., and Helms, V. (2002) Proton shuttle in green fluorescent protein studied by dynamic simulations, *Proc. Natl. Acad. Sci. USA 99*, 2778-2781.
67. Bondar, A. N., Elstner, M., Suhai, S., Smith, J. C., and Fischer, S. (2004) Mechanism of primary proton transfer in bacteriorhodopsin, *Structure 12*, 1281-1288.
68. Pomes, R., and Roux, B. (1996) Structure and dynamics of a proton wire: a theoretical study of H+ translocation along the single-file water chain in the gramicidin A channel, *Biophys. J. 71*, 19-39.
69. Pomes, R., and Roux, B. (1998) Free Energy Profiles for H+ Conduction along Hydrogen-Bonded Chains of Water Molecules, *Biophys. J. 75*, 33-40.
70. Xu, J. C., and Voth, G. A. (2005) Computer simulation of explicit proton translocation in cytochrome c oxidase: The D-pathway, *Proc. Natl. Acad. Sci. USA 102*, 6795-6800.
71. Braun-Sand, S., Strajbl, M., and Warshel, A. (2004) Studies of proton translocations in biological systems: Simulating proton transport in carbonic anhydrase by EVB-based models, *Biophys. J. 87*, 2221-2239.
72. Olkhova, E., Helms, V., and Michel, H. (2005) Titration behavior of residues at the entrance of the D-pathway of cytochrome c oxidase from Paracoccus denitrificans investigated by continuum electrostatic calculations, *Biophys. J. 89*, 2324-2331.

73. Chen, H., Wu, Y., and Voth, G. A. (2006) Origins of Proton Transport Behavior from Selectivity Domain Mutations of the Aquaporin-1 Channel, *Biophys. J. 90*, L73-75.
74. Mathias, G., and Marx, D. (2007) Structures and spectral signatures of protonated water networks in bacteriorhodopsin, *Proc. Natl. Acad. Sci. USA 104*, 6980-6985.
75. Lill, M. A., Hutter, M. C., and Helms, V. (2000) Accounting for environmental effects in ab initio calculations of proton transfer barriers, *J. Phys. Chem. A 104*, 8283-8289.
76. Lill, M. A., and Helms, V. (2001) Reaction rates for proton transfer over small barriers and connection to transition state theory, *J. Chem. Phys. 115*, 7985-7992.
77. Lill, M. A., and Helms, V. (2001) Molecular dynamics simulation of proton transport with quantum mechanically derived proton hopping rates (Q-HOP MD), *J. Chem. Phys. 115*, 7993-8005.
78. Lill, M. A., and Helms, V. (2001) Compact parameter set for fast estimation of proton transfer rates, *J. Chem. Phys. 114*, 1125-1132.
79. de Groot, B. L., Frigato, T., Helms, V., and Grubmuller, H. (2003) The mechanism of proton exclusion in the aquaporin-1 water channel, *J. Mol. Biol. 333*, 279-293.
80. Gu, W., Frigato, T., Straatsma, T. P., and Helms, V. (2007) Dynamic Protonation Equilibrium of Solvated Acetic Acid, *Angew. Chem. Int. Ed. 46*, 2939-2943.
81. Herzog, E., Frigato, T., Helms, V., and Lancaster, C. R. D. (2006) Energy Barriers of Proton Transfer Reactions between Amino Acid Side Chain Analogs and Water from ab initio Calculations., *J. Comput. Chem. 27*, 1534-1547.
82. Kay, B. K., Williamson, M. P., and Sudol, P. (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains, *Faseb J. 14*, 231-241.
83. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.* 261, 470 – 489

Chapter 2

Alternative Binding Modes of Proline-rich Peptides Binding to the GYF Domain

(published in *Biochemistry*, **44**, 6404-6415 (2005))

2.1 Summary

Recognition of proline-rich sequences plays an important role for the assembly of multi-protein complexes during the course of eukaryotic signal transduction and is mediated by a set of protein folds that share characteristic features. The GYF (glycine-tyrosine-phenylalanine) domain is known as a member of the super-family of recognition domains for proline-rich sequences. Recent studies on the complexation of the CD2BP2-GYF domain with CD2 peptides showed that the peptide adopts an extended conformation and forms a polyproline type II helix involving residues Pro4 – Pro7 [Freund et al. (2002) *EMBO J. 21*, 5985-5995.]. R/K/GxxPPGxR/K is the key signature for the peptides that bind to the GYF domain [Kofler et al. (2004) *J. Biol. Chem. 279*, 28292-28297]. In our combined theoretical and experimental study, we show that the peptides adopt a polyproline II helical conformation in the unbound form as well as in the complex. By molecular dynamics simulations we identify a novel binding mode for the G8W mutant and the wild-type peptide (shifted by one proline in register). In contrast, the conformation of the peptide mutant H9M remains close to the experimentally derived wild-type GYF-peptide complex. Possible functional implications of this altered conformation of the bound ligand are discussed in the light of our experimental and theoretical results.

2.2 Introduction

Intracellular protein domains recognizing proline-rich sequences (PRS) play a pivotal role in biological processes that require the coordinated assembly of multi-protein complexes (*1*). In vertebrate genomes, PRS are predicted to be among the most abundantly expressed amino acid sequence motifs (*2*) and this corresponds to an increasing number of proteins that acquired PRS-recognition domains during the course of evolution (*3*).

Up to now, the super-family of proline-rich sequence recognition domains consists of profilin (*4*), the SH3 (*5, 6*), the WW (*7*), the EVH1 (*8*), the GYF (*9, 10*), the UEV (*11, 12*) and probably the ligand binding domain of prolyl-4-hydroxylase (*13*). For each of these domains a set of conserved aromatic amino acid residues is important for peptide binding. Within the GYF domain the glycine-tyrosine-phenylalanine tripeptide is part of a bulge-helix-bulge motif that contains several aromatic amino acid side-chains that are essential for the binding of the CD2 cytoplasmic domain. The recently solved NMR structure of the CD2BP2-GYF domain in complex with the CD2 peptide SHRPPPPGHRV (*14*) showed that the peptide adopts an extended conformation and forms a left-handed polyproline type II (PPII) helix involving residues Pro4 – Pro7 (*14, 15*) (see Figure 1). The binding surface of the GYF domain accommodates Pro6 and Pro7 of the ligand and is defined primarily by the aromatic residues Tyr6, Trp8, Tyr17, Tyr20, Trp28, Tyr33 and Phe34 of the GYF domain (see Figure 1).

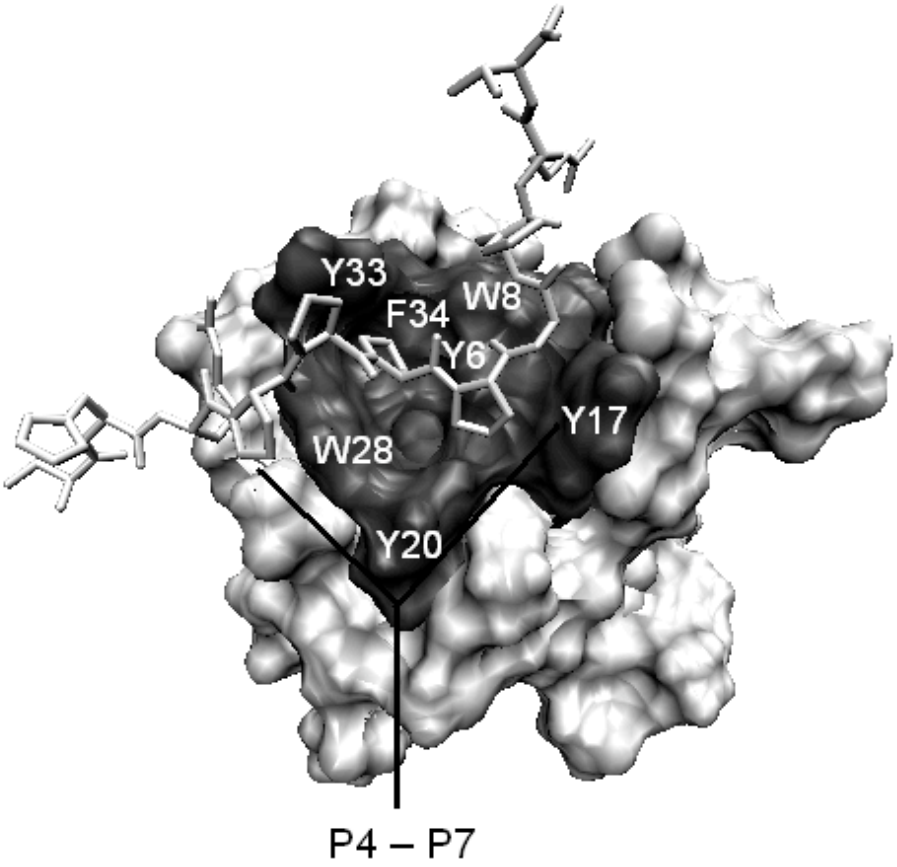


Figure 1: NMR structure of GYF domain with wild-type peptide. The GYF domain is represented by its molecular surface; the peptide atoms are drawn as sticks. Residues forming the binding pocket are coloured in dark grey and labelled by their one-letter codes and sequence numbers. The VMD (*60*) package was used to generate this picture.

Characterizing the conformational changes for both interaction partners is essential for understanding the mechanism of the peptide binding to the GYF domain.

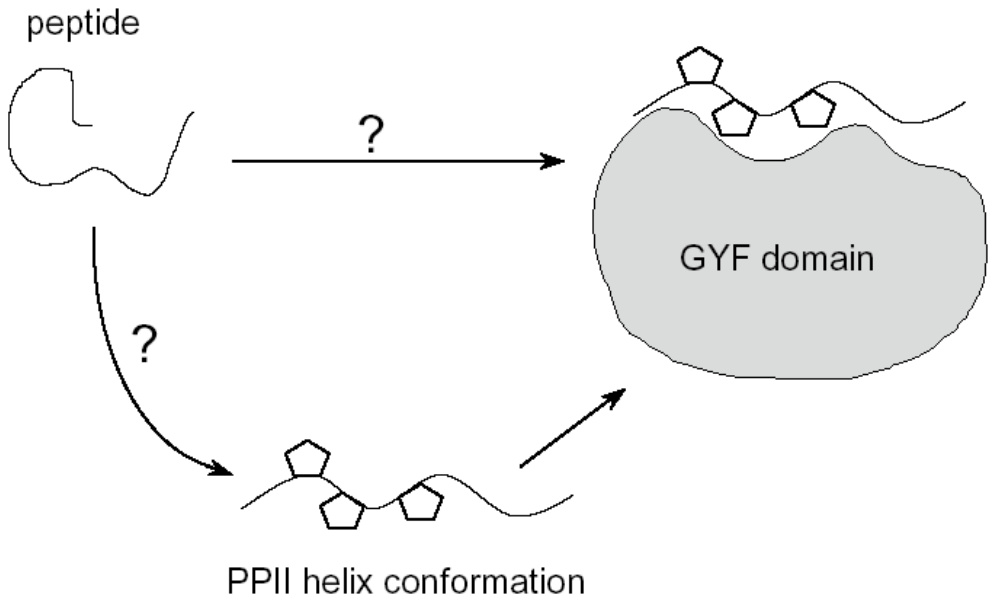
On one hand, it has been recently recognized that many proteins contain long disordered segments in their functional states under usual physiological conditions (*16-20*), e.g. most of the polypeptide hormones are conformationally disordered in aqueous solution and fold upon binding to their receptors (*20*). Unstructured segments within large proteins provide ideal scaffolds for the interaction with several different targets and thereby help to assemble multi-protein complexes (*16-20*). On the other hand, it has been shown by many experimental and theoretical studies that certain peptides, including proline-rich sequences, adopt preferred conformations in solution (*1, 21-31*). Therefore, it is a matter of ongoing discussion whether the PPII helix is such a preferred conformation of certain peptide sequences (*21, 25, 27-29, 32-39*). In the CD2 polyproline peptide-GYF complex, the central part of the peptide adopts a PPII helical conformation. A mechanistic description of the binding event has to distinguish whether the PPII helix conformation is preformed in the unbound peptides and binding to the GYF domain takes place in a "lock and key" mode or whether folding and binding occurs in parallel, corresponding to an 'induced fit' model (Scheme 1). Further it should clarify which conformational changes take place in the protein and the peptide and how these changes contribute to the stabilization of the complex. So far, our previous study has identified the key binding motif of the peptide as R/K/GXXPPGXR/K (*40*). The structural importance of peptide Gly8 for interaction with the GYF domain has also been analyzed (*14*): A glycine in this position terminates the PPII helix conformation and prevents hindrance between C-terminus of the peptide and the domain. A G8X mutation resulted in loss of binding for most residues during systematic mutagenetic studies (see Table 1). Surprisingly, the peptide still binds to the domain upon a G8W mutation (*40*), although, based on the wild-type structure, a G8W substitution would result in a clash between the tryptophane and the GYF domain. Considering the large structural differences between glycine and tryptophan, a different binding mode for this peptide can be assumed. It has been shown by NMR experiments (*41*) that SH3 domains can bind proline rich ligands in two orientations, due to the pseudo-symmetry of the PPII helix. These findings raised the question whether such a scenario is also true for the GYF domain.

In the present work, we carried out theoretical calculations to address the problem of the conformational state of unbound peptides. Molecular dynamics (MD) simulations starting from different initial conformations and at different temperatures indicated that all studied peptides adopt the PPII helical conformation in the unbound state. For wild type and G8W mutant peptide we combined NMR experiments with theoretical calculations and identified a novel binding mode (register shifted by one proline), while the control peptide mutant H9M remains close to the experimental GYF-domain wild-type peptide complex conformation (*14*). Possible functional implications of this altered conformation of the bound ligand are discussed in the light of our experimental and theoretical results.

2.3 Materials and Methods

2.3.1 Protein production and NMR analysis

The GYF domain of human CD2BP2 comprising amino acids 280–342 was cloned and expressed as described elsewhere (*40*). The NMR experiments were performed at



Scheme 1

296 K using a Bruker DRX600 instrument equipped with a standard triple-resonance probe. Data processing and analysis were carried out using the XWINNMR (Bruker) software package and the program Sparky (*42*). In the NMR experiments, increasing amounts of the synthetic peptide of sequence NH_2-SHRPPPPGHRV–COOH or NH_2-SHRPPPPWHRV–COOH were added to a 0.2 mM sample of the ^{15}N labeled GYF domain up to a final concentration of 1.8 mM. HSQC spectra were recorded and the changes of the assigned nitrogen and hydrogen chemical shifts were combined as follows: $[(\Delta^1H\_cs)^2 + (\Delta^{15}N\_cs)^2]^{1/2}$, where Δ^1H_cs is the chemical shift change for ^1H atoms in units of 0.1 p.p.m., and $\Delta^{15}N\_cs$ is the chemical shift change for ^{15}N atoms in units of 0.5 p.p.m. The sum of these weighted geometrical differences of the chemical shifts were plotted against the peptide concentrations for the titration experiments. The resonances of residue F34 and of the W8 side chain were excluded due to line broadening preventing the identification of the corresponding resonances at various ligand concentrations. The dissociation constants were calculated using the program Microcal™ Origin™. For comparison of the two binding epitopes, the weighted geometrical differences of the chemical shifts for all assigned residues upon addition of 1.8 mM ligand are shown in a histogram.

2.3.2 Peptide substitution analysis

WT	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
S	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
H	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
R						+			+						+					
P	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
P	+					+							+							
P													+							
P						+							+							
G						+									+				+	
H	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
R									+						+					
V	+				+	+	+	+	+	+	+	+	+	+	+	+	+	+		+

Table 1 Favorable mutations of the peptides binding to GYF domain. Data are taken from Kofler et. al. 2004 (ref 40), mutations favorable for the binding are marked "+". The first column labeled 'WT' contains the sequence of wild-type peptide. The later columns contain the results from single-amino acid mutation experiments. Amino acid residues are listed with their one letter code.

Single substitutions of SHRPPPPWHRV and a set of different proline rich peptides were generated by semiautomated spot synthesis (*43, 44*) (Abimed; Software LISA, Jerini AG) on Whatman 50 cellulose membranes as described (*45*). Membranes were probed with GST fusion protein as described elsewhere (*46*). Briefly, the membranes were incubated with GST-GYF (CD2BP2; 40 μg/ml) over night. After washing, bound GST fusion protein was detected with rabbit polyclonal anti-GST antibody (Z-5, Santa Cruz) and horseradish peroxidase coupled anti-rabbit IgG antibodies (Rockland). An enhanced chemiluminescence substrate (SuperSignal West Pico, Pierce Illinois) on a LumiImagerTM (Boehringer Mannheim GmbH) was used for detection.

2.3.3 Molecular dynamics simulations

Peptides. To characterize the conformational ensembles of the unbound solvated peptides, a set of MD simulations of the wild-type (SHRPPPPGHRV) and the mutated peptides (SHRPPPPWHRV and SHRPPPPGMRV) were carried out using the GROMACS3.14 package (*47*) applying the OPLSAA force field (*48*). In some cases the starting structures were taken from the complex of the wild-type peptide with the GYF domain (PDB entry 1L2Z (*14*)). Three MD simulations of the wild-type peptide were performed: 1) start from the wild-type NMR structure at 300 K

temperature (WT); 2) start from a modeled extended structure at 300 K temperature (WTE); 3) start from the NMR structure at a temperature of 500 K (WTHT). The extended conformation in the WTE simulation was generated with dihedral angles of backbone of 135 degrees (N-CA-C-N), 180 degrees (CA-C-N-CA) and –135 degrees (C-N-CA-C). MD simulations of the mutated peptides (G8W for SHRPPPPWHRV and H9M for SHRPPPPGMRV) were performed only at 500 K. Mutated residues (tryptophan in the G8W simulation and methionine in the H9M simulation) were modeled using the TINKER package (*49*) based on the NMR structure of the wild-type peptide in the complex. The peptides were solvated in cubic boxes, using TIP3P water molecules (*50*), with an initial minimum distance of at least 14 Å between the boundaries of the box and the nearest solute atom. All coordinate sets were first minimized by 500 steps of steepest-descent energy minimization. The solvent and protein atoms were then relaxed during a 100 ps MD simulation with all non-hydrogen atoms of the NMR structure restrained to their coordinates in the PDB structure. Then plain MD simulations (20 ns for WT, WTE and WTHT and 40 ns for G8W and H9M) were carried out without any restraints. The LINCS procedure (*51*) was applied to constrain all bond lengths. The time step of the simulation was set to 2 fs. A 9 Å cutoff was used for the short-range non-bonded interactions and the lists of non-bonded pairs were updated every 10 steps. The Particle Mesh Ewald (PME) method (*52*) with a grid size of 1.2 Å was used to calculate long-range electrostatic interactions. Temperature and pressure were maintained by weak coupling to an external bath in the simulations (*53*). Cluster analysis was carried out after the simulations using the "full linkage" algorithm implemented in the GROMACS3.14 package (*47*): a structure is added to an existing cluster when its distance to any element of the cluster is less than the given cutoff. Main-chain atoms and C_β atoms were selected for calculating the RMSD matrix. The RMSD cutoff was set to 1.5Å.

Complexes. The dominant conformations in the cluster analysis of each simulation of the unbound mutated peptides G8W and H9M were superimposed on the wild-type peptide in the structure of the complex using all main-chain atoms (backbone, H and O) and the C_β atoms of the HRPPPP segment for the alignment. Then 500 steps of steepest-descent energy minimization were applied to remove unfavorable interactions. The two optimized mutated complexes were used as starting structures in the simulations of the mutant complexes. Similar procedures as for the peptide simulations were used to carry out two 30 ns long MD simulations of the modeled complexes at 300 K (G8W_GYF for the peptide SHRPPPPWHRV and the GYF domain, and H9M_GYF for the peptide SHRPPPPGMRV and the GYF domain). To investigate the binding mode of the G8X mutations more systematically, we used the package FlexX (*54*) to dock the G8W, G8R, G8Y and G8K mutants to the GYF domain followed by subsequent MD simulations. Pro7 of the peptide was chosen as a seed in the docking while the complex conformations with the best docking score were chosen as the starting structures of the simulations. MD simulation of the wild-type complex was also carried out as a control run (WT_GYF). Details about all simulations (starting structure, temperature, and length) reported here are summarized in Table 2.

Simulation	System	Starting Structure	T	Length
WT	Wild type peptide	NMR structure	300 K	20 ns
WTE	Wild type peptide	Modeled extended structure	300 K	20 ns
WTHT	Wild type peptide	NMR structure	500 K	20 ns
G8W	G8W mutant peptide	Modeled from NMR	500 K	40 ns
H9M	H9M mutant peptide	Modeled from NMR	500 K	40 ns
WT_GYF	Wild type + GYF	NMR structure	300 K	30 ns
G8W_GYF	G8W mutant + GYF	Modeled from NMR and simulation	300 K	30 ns
H9M_GYF	H9M mutant + GYF	Modeled from NMR and simulation	300 K	30 ns
G8W_DOCK	G8W mutant + GYF	Docking based on NMR	300 K	30 ns
G8R_DOCK	G8R mutant + GYF	Docking based on NMR	300 K	30 ns
G8Y_DOCK	G8Y mutant + GYF	Docking based on NMR	300 K	30 ns
G8K_DOCK	G8K mutant + GYF	Docking based on NMR	300 K	30 ns
ALT_GYF	Wild type + GYF	Modeled from NMR and simulation	300 K	5*20 ns

Table 2 Summary of all the simulations

2.4 Results

2.4.1 Solvent conformation of the unbound peptide

As described in the Methods section, the molecular dynamics simulations for the wild-type peptide were started from different starting conformations of the peptide (a PPII helical conformation taken from the NMR complex and a modeled extended starting conformation) at two different temperatures (300k and 500k). Figure 2 shows the evolution of backbone dihedral angles (Φ, Ψ) during the 20 ns long simulations. In the WTE run, the formation of the PPII helical conformation occurred after only a few ps. Therefore, this initial conformation cannot be resolved in Figure 2a. In all three simulations of the wild-type peptide (shown in Figure 2a), the backbone dihedral angles of residues His2-Pro7 merely fluctuated around the ideal value for PPII helix: Φ = -78º, Ψ = 146º (*55*). The simulation at high temperature (WTHT) shows only slight shifts, where the segment Gly8-Val11 contributes to most of the fluctuation of the backbone conformation. Similar results were also found in the simulations of the mutated peptides G8W and H9M (shown in Figure 2b).

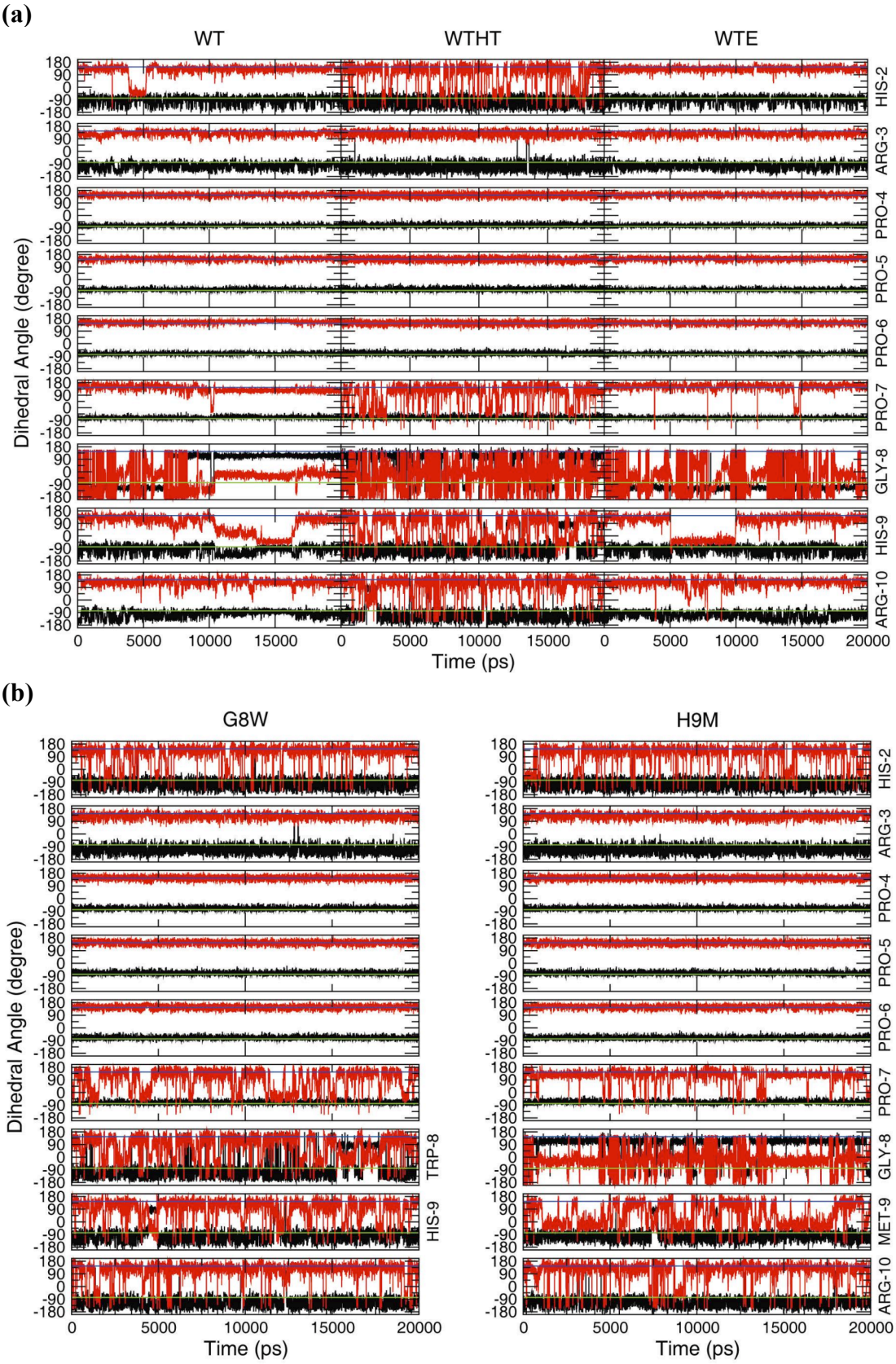


Figure 2: Evolution of the backbone dihedral angles (black: Phi angles; red: Psi angles) during the simulation of the wild-type peptide (a) and the mutant peptide (b). Ideal values of the dihedral angles are shown in solid lines (blue: Phi angles; green: Psi angles).

After this initial comparison between the simulations of wild-type and mutated peptide, the G8W and H9M simulations were extended to 40 ns to improve the sampling of the conformational space of the mutated peptides and to allow for identification of possible interesting peptide conformations that could be used for the modeling of the mutated complexes. We used cluster analysis to summarize the sampling during the simulation of G8W and H9M. Interestingly, only one dominant cluster was found in each simulation: The largest cluster covers 77% and 86% of the trajectory in the G8W and H9M simulation, respectively. A similar observation was made for the simulations of the wild-type peptide: only one cluster was found in the analysis, which covers almost the whole trajectory. These results agree with the dihedral angle analysis that the backbones of the mutated and wild-type peptides are quite stable during the simulation.

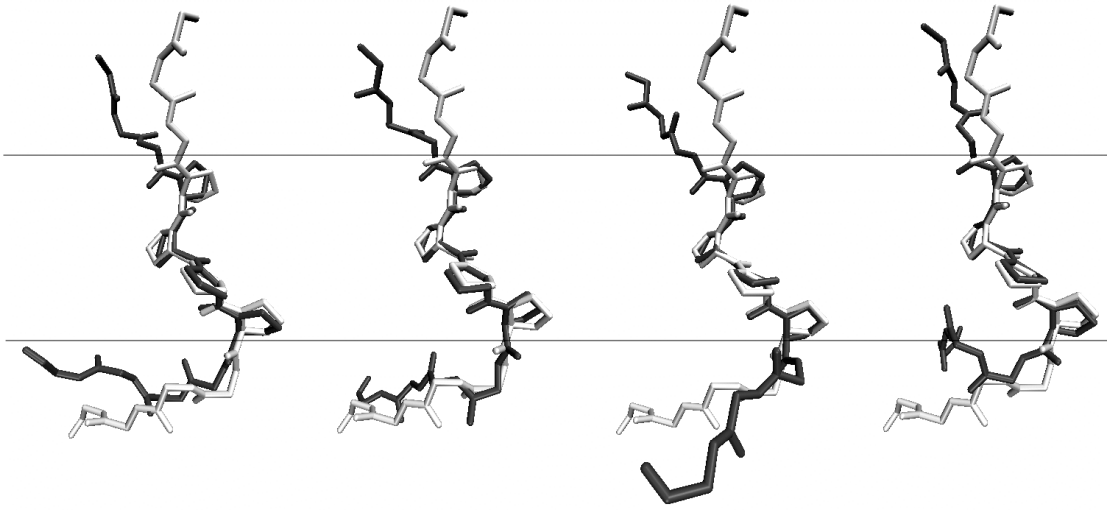


Figure 3: Superposition of the representative conformations of simulations of unbound peptides (from left to right: WT, WTE, G8W and H9M) onto the bound peptide in the NMR structure. Representative conformations are colored in black while the bound peptide in the NMR structure is shown in grey.

Figure 3 shows a superposition of representative conformations of each simulation (WT, WTE, G8W and H9M) onto the bound peptide of the NMR structure. It is clearly visible that the PPII helix conformation is adopted in all cases. The conformations of Pro4-Pro7 overlap very well with each other, and deviations only appear at the two termini of the peptides. The stable PPII helix conformation found in all simulations indicates that all three peptides are able to adopt a PPII helix conformation in the unbound state and is reflected by the occurrence of one dominant cluster. This is in agreement with previous theoretical and experimental studies on polyproline peptides (*1, 21, 23-26*). Rucker and Creamer explained the bias of polypeptide folding into the PPII helix as an energetically favorable option: all backbone polar groups are well-solvated in this conformation in water, thus compensating for the lack of intramolecular hydrogen bonds (*28*).

2.4.2 Binding analysis of the GYF domain to the mutated and wild-type peptides

Binding analysis and combined chemical shift changes were measured by NMR experiments for the G8W peptide as well as for the wild-type peptide binding to CD2BP2 GYF domain. The results are shown in Figure 4a-4c. The spectra, and therefore the chemical shift changes of the GYF domain in complex with wild-type peptide (SHRPPPPGHRV) and the mutated peptide G8W (SHRPPPPWHRV) are very similar. Since the chemical shift is a very sensitive measure of the chemical environment, the precise overlap for almost all resonances except Trp8 in both spectra (Figure 4b) surprisingly demonstrates that the binding surface of the GYF domain is very similar for the two peptides. The pattern of chemical shift changes (Figure 4c) reveals the binding face for the polyproline peptide on the GYF domain. Most of the strongly shifted resonances belong to residues that are highly conserved among putative GYF domains (*14*) and are almost identical for the two peptides. However, the G8W peptide binds with slightly lower affinity. Assuming a two-state binding model for the peptides, apparent K_D values of 220 ± 30 μM and 290 ± 20 μM were determined by NMR for the wild-type and the mutant peptide, respectively. Large differences of the chemical shift change were found for the backbone NH of W8 (Figure 4b). This result illustrates that conformational changes due to the G8W mutation happen near this residue, while for other residues, the chemical shift changes are quite similar between the GYF domains binding to the wild-type peptide and the G8W mutant.

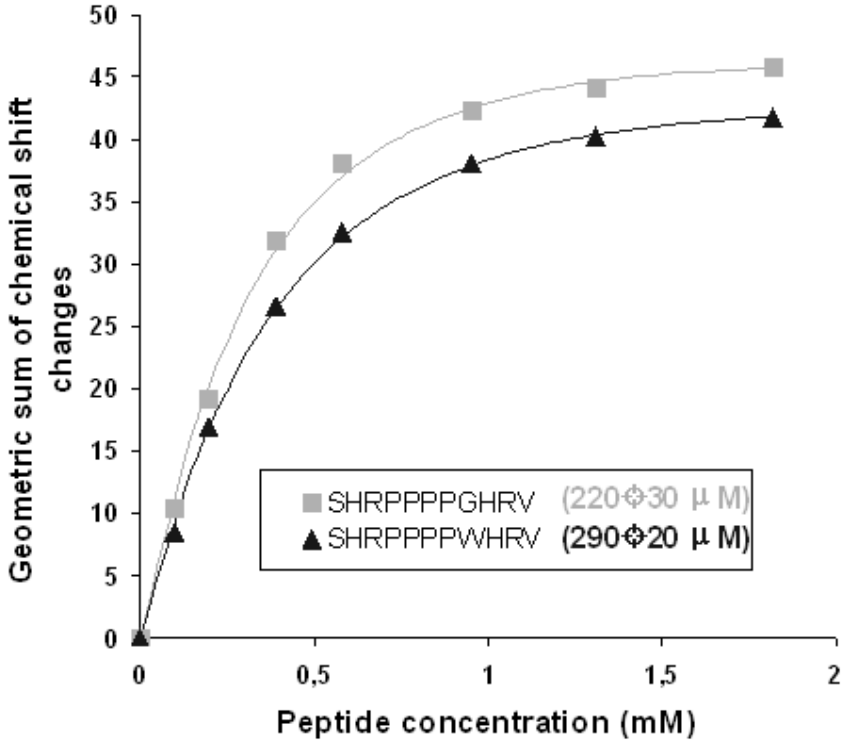
2.4.3 Structure of the complex with the mutated peptide

To investigate the conformational changes of the complex, which are due to single mutations of peptide residues, we first modeled both mutant complexes: Gly8 to Trp (G8W_GYF) and His9 to Met (H9M_GYF). In the G8W_GYF complex, the mutated residue Gly8 is very important for the binding of the peptide, while in the second case, the mutated residue His9 is not crucial for the binding affinity as shown by systematic mutational analysis (Table 1). Considering that the mutated peptides only adopt one dominant conformation for their polyproline regions and that the mutations were introduced only near the C-terminus, we superimposed the mutated peptides on the wild-type peptide using the PPII helix and the N-terminus segment thereby modeling the initial structure of the mutated complexes based on the superposition.

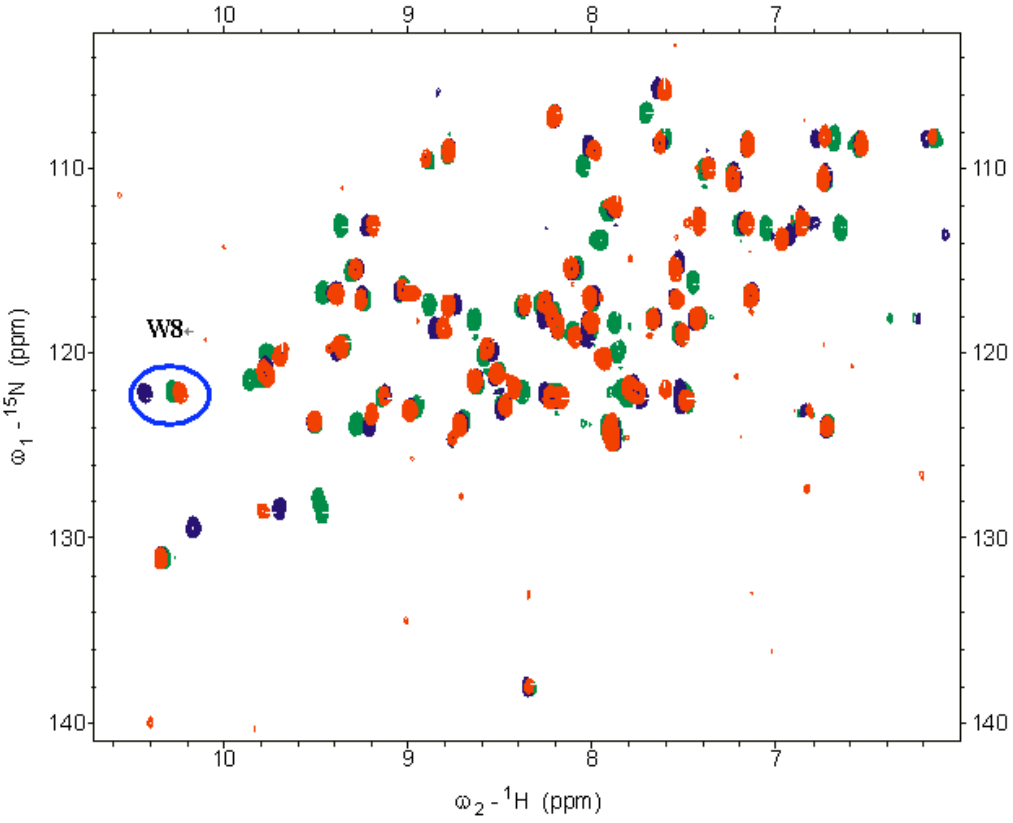
Overview of the simulations. The two modeled mutant complexes were first optimized by 500 steps of steepest-descent energy minimization. However, energy minimization is not enough to provide a complete picture of the properties and stabilities of the predicted structures and should be complemented by unconstrained MD simulations. The root mean square deviations (RMSD) with respect to the starting and final coordinate sets of the C_α atoms are very stable and center around 2 Å in the simulations of the wild-type (WT_GYF) and the H9M mutant (H9M_GYF) complex. For the G8W_GYF complex however, the RMSD of the simulation are significantly smaller with respect to the final coordinates compared to the initial coordinates (2 Å and 3-3.5 Å, respectively). This indicates that the modeled starting conformation is not stable and some conformational changes occurred during the

simulation while the final structure is more stable and is closer to the "central" conformations of the simulations.

(a)



(b)



(c)

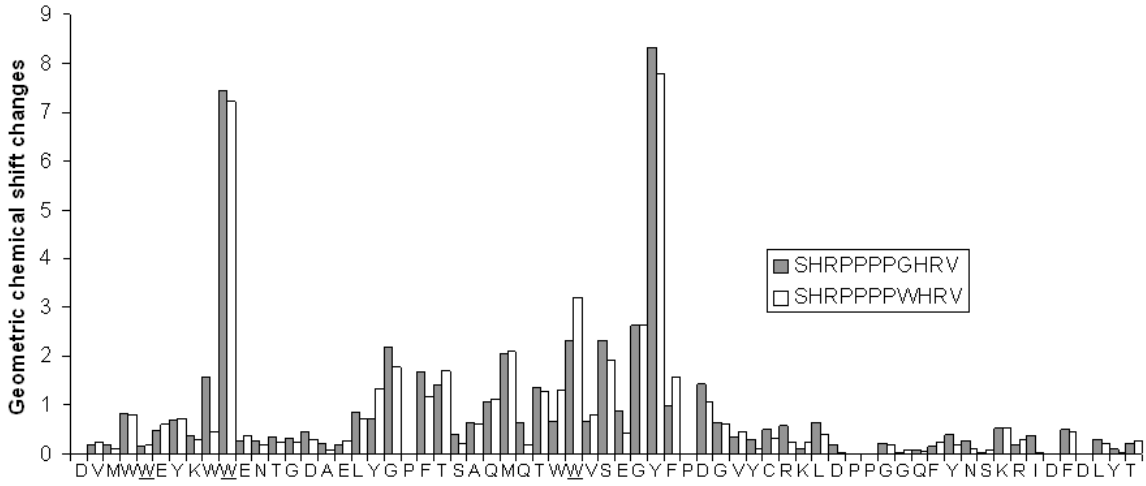


Figure 4: Binding analysis of the CD2BP2-GYF domain to the peptide SHRPPPPWHRV in comparison to the wild-type peptide SHRPPPPGHRV by NMR. (a) The sum of the weighted geometrical differences of the chemical shifts (Geometric sum of chemical shift changes) for assigned peaks, which could be identified at all applied peptide concentrations is plotted against the concentration of the peptide. (b) Mapping of the binding site of SHRPPPPGHRV and SHRPPPPWHRV peptides onto the CD2BP2-GYF domain. Overlay of HSQC spectra of GYF domain alone (green) and GYF-domain in the presence of a 10-fold excess of the wild-type peptide SHRPPPPGHRV (blue) or the mutant peptide SHRPPPPWHRV (red), respectively. A quantitative analysis of the chemical shift changes of each residue is presented as histogram (c). The weighted geometrical differences of the chemical shifts for each assigned residue upon addition of a 10-fold excess of peptide are plotted against the corresponding residue. Prolines are depicted for completeness. The weighted geometrical differences of the chemical shifts of tryptophan side chains are indicated by W.

A more systematical way of characterizing the conformations is the before mentioned cluster analysis. Main-chain atoms and C_β atoms were selected for calculating the RMSD matrix. The RMSD cutoff was set to 1.0 Å. This cutoff is smaller than that used for the MD simulations of the peptides (1.5 Å). Nevertheless, a few dominant clusters cover most of the trajectory: in the G8W_GYF simulation, the two largest clusters cover 81% of the trajectory (46% and 35%, respectively) and the remaining 18 clusters share the remaining 19%, while in the H9M_GYF simulation only one large cluster was found, which covers almost 99% of the trajectory. In the superimposed model of the G8W_GYF complex, a new pocket is opened by Trp8, Glu9, Tyr17, and Phe20 of the domain. It seems that this pocket accommodates the large side-chain of Trp8 of the peptide and therefore avoids clashes. However, this binding mode was not stable in the simulation and the side-chain of peptide residue Trp8 finally moved out of this pocket after 12 ns, pointing towards the solvent where it finally reached an equilibrated state. This change is also reflected by the cluster analysis: two large clusters were found in the simulation, each representing one state of the Trp8 side-chain. The cluster in which the side-chain of peptide Trp8 stays in the pocket covers most of the trajectory between 0-12 ns (35% of total and 87% of 0-

12 ns), while the other one in which Trp8 moved out covers most of the remaining part (46% of total and 77% of 12-30 ns).

Docking experiments. To investigate the binding mode of the G8X mutations more systematically, we docked G8W, G8R, G8Y and G8K mutant to the GYF domain using the FlexX program for flexible ligand docking and then carried out MD simulations (G8X_DOCK). The G8W and the G8R mutations are the only mutations that were experimentally confirmed to be favorable in this position (Table 1). The G8Y and the G8K mutants were chosen to mimic G8W and G8R as control cases. After docking, the side-chains of the mutated residue (G8X) pointed to the solvent in all cases implicating that the prolines are shifted by one position. This "shift in register" is in agreement with the experimental structure of the complex (used for the docking) that seems not to allow larger side chains at the position of Gly8. In the G8W_DOCK simulation, a contact was found between the side-chain of Trp8 of the peptide and the side-chain of Trp8 of the GYF domain. The distance between the center of mass of the two side-chains was 5.9 ± 1.1 Å during the 30 ns MD simulation. The same contact was also found for the G8R_DOCK simulation, where the distance between the center of mass of Arg8 (peptide) and Trp8 (GYF domain) was 4.3 ± 0.5 Å during the simulation. For the G8K_DOCK simulation, this contact is formed only during the first 5 ns of the simulation and is finally lost in the remaining simulation (5 ns – 30 ns). The distance between the center of mass of the corresponding residues shifted from 4.6 ± 0.8 Å (0 ns – 5 ns) to 11.2 ± 1.3 Å (5 ns – 30 ns). In the G8Y_DOCK simulation, no contact between Tyr8 of the peptide and any residues in the GYF domain was found during the entire simulations. The average distance between the center of mass of Tyr8 (peptide) and Trp8 (GYF domain) was 12.9 ± 1.5 Å.

2.5 Discussion

2.5.1 Preformation of the PPII helix

For many proteins with unstructured segments, the coupling of binding and folding is favorable according to the binding free energy: the entropic penalty associated with the folding transition is counterbalanced by a large enthalpy of binding (*16, 56*). In those cases, the folding upon binding acts as a fine controller of the thermodynamics balance. In contrast, the polyproline peptides in our study are already folded into a PPII helix conformation in the unbound state and bind constitutively to the GYF domain. This binding mode is entropically more favorable than binding of unstructured peptides. The rigid PPII helix conformation of the unbound peptides studied is intrinsically stable in solution and is also favorable for its specific binding motif. Hilser and colleagues studied binding of the polyproline Sos peptide to the Sem-5 SH3 domain (*21*). They found that the PPII bias of unstructured peptides is driven by a favorable and significant enthalpy (ΔH) of –1.7 kcal mol^{-1} $residue^{-1}$, which is partially offset by an unfavorable entropy ($T\Delta S$) of –0.7 kcal mol^{-1} $residue^{-1}$, relative to the ensemble of disordered conformation of the molecule. A similar example is the c-Myb oncoprotein, which folds into an α-helical conformation both complexed and uncomplexed with its target protein (*56*). Remarkably, binding of c-

Myb to its target (residue 586-672 of CREB binding protein) is entropically favored (ΔS = +7.5 cal mol^{-1} K^{-1}) while its favorable enthalpy change is small (ΔH = -4.1 kcal mol^{-1} K^{-1}) (*16, 56*).

It has been proposed by Dyson and Wright that unstructured proteins provide a large flexibility of binding reactions because they may adopt various structures upon binding to different partners (*20*). On the other hand, as examplified here for the GYF domain-ligand pair, the preformation of a peptide conformation might be well suited to guarantee the rapid formation of specific peptide-protein complexes within the dynamic settings of signal transduction.

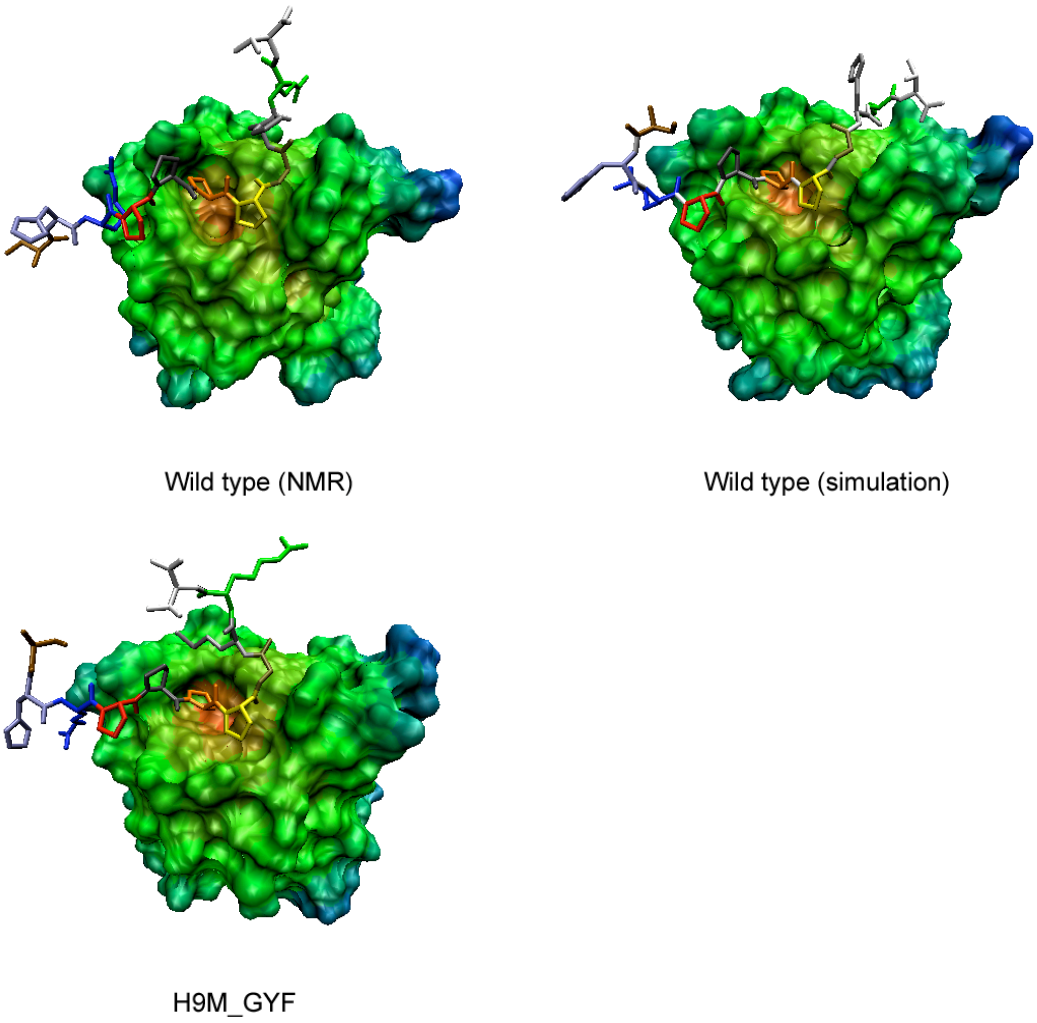
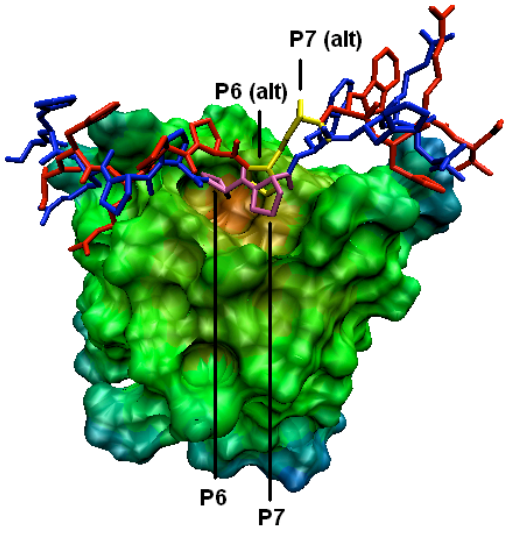


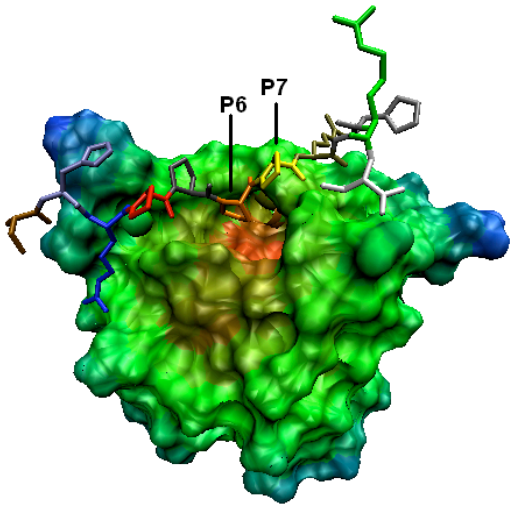
Figure 5: Comparison of the binding interfaces of the GYF domain (NMR and simulation) for the wild-type complex (above) and of the H9M mutant (below). The GYF domain is represented by its molecular surface and coloured by position (from orange to deep blue: completely buried to completely exposed); the peptide atoms are drawn as sticks and coloured according to their appearance in sequence.

2.5.2 Analysis of the binding modes

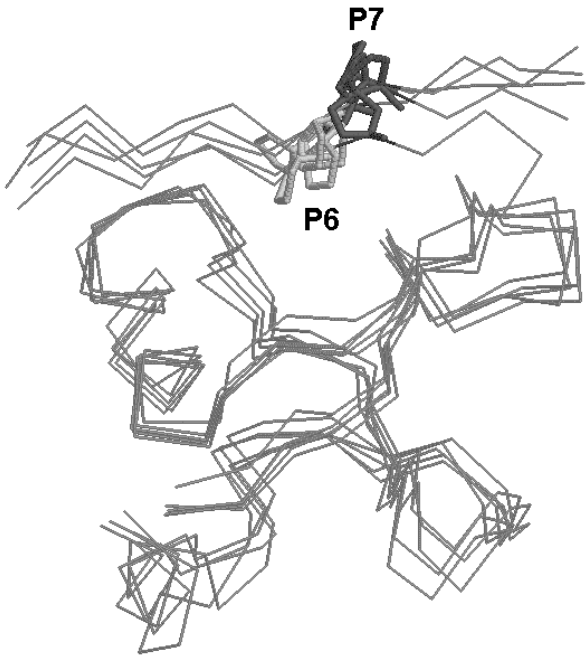
(a)



(b)



(c)



(d)

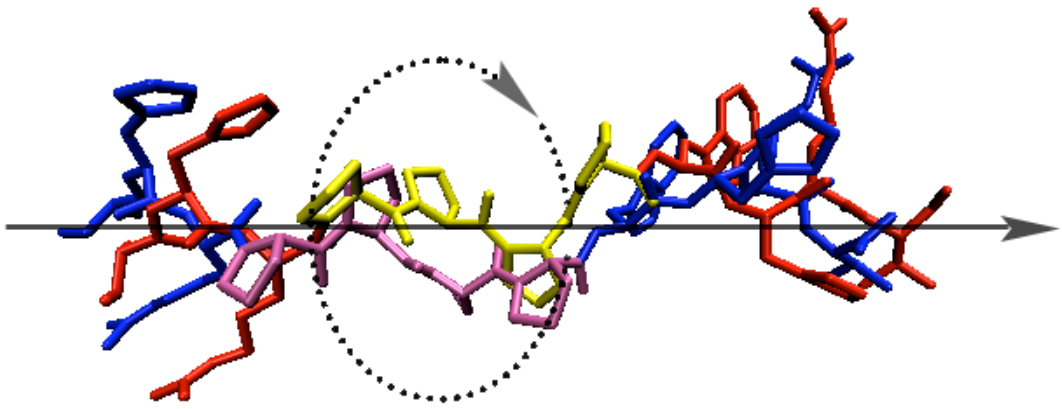


Figure 6: (a) Superposition of the two binding modes found in the simulation of the G8W mutant complex (starting from the docking results). The two conformations of the peptide are drawn as sticks (blue: mode 1, red: mode 2, pink: Pro6 and Pro7 in mode 1, yellow: Pro6 and Pro7 in mode 2). (b) Binding mode of the G8R mutant complex (representative conformation of the simulation). The peptide atoms are represented by sticks and coloured according to their sequence number. In (a) and (b), the GYF domain is represented by its molecular surface and coloured by position (from orange to deep blue: completely buried to completely exposed) and Pro6 and Pro7 are labelled by their one-letter codes and sequence numbers. Mode 2 is labelled as "(alt)". (c) Superposition of the representative conformations of the five simulations of wild type GYF complex starting from the alternative binding mode. Pro6 and Pro7 are represented by sticks and are labelled by their one-letter codes and sequence numbers. Pro6 is coloured in light grey and Pro7 is coloured in dark grey. (d) The translation and rotation motions of the peptide between the two binding modes (blue: mode 1, red: mode 2, pink: Pro4 to Pro7 in mode 1, yellow: Pro4 to Pro7 in mode 2). For Pro4 to Pro7 a rotation is the principle component of motion, while for other residues in the peptide a translation is the principle component of motion.

H9M_GYF agrees with the wild-type. The results from the cluster analysis of the simulations as well as the NMR structures were used for comparing the binding modes for different mutations. Figure 5 shows the binding interfaces as well as the peptide of the wild-type complex (NMR and simulation) and that of the H9M mutant with the GYF domain The hydrophobic pocket formed by Trp8, Tyr17, Phe20, Trp28 and Tyr33 of the GYF domain is structurally maintained to accommodate Pro6 and Pro7 of the peptide in the simulations of WT_GYF and H9M_GYF, in agreement with the NMR derived structure. Arg3 and Arg10 of the peptide stay close to Glu31 and Glu9 of the domain. We conclude that favorable hydrophobic interactions between Pro6, Pro7 and the pocket, together with the electrostatic attraction between the positively charged residues Arg3 and Arg10 of the peptide and the negatively charged residues Glu31 and Glu9 of the domain play a central role for binding. These observations are consistent with the results from the substitution analysis — any mutation among these residues induces an unbinding of the peptide and the only tolerated substitutions are to glycine or lysine, and to lysine for Arg3 and Arg10, respectively (see Table 1). Other important interactions present in all simulations as well as in the NMR experiments are the hydrogen bonds between the backbone oxygens of peptide Pro4 and Pro7 and the side-chain H_ε in the domain. However, these hydrogen-bonding interactions do not bring any specificity for Pro4 in that position because most substitutions are tolerated. (see Table 1). This can be explained by the fact that the H_α of Pro4 points into solution and no replacement should cause clashes with other residues or influence the formation of this hydrogen bond. For Pro7, the specificity is the consequence of the hydrophobic interaction with the binding pocket of the GYF domain.

In the WT_GYF and H9M_GYF simulations, the binding interfaces do not show significant differences with respect to the NMR structure. The small RMSD in these two simulations also indicate that the wild-type peptide and the H9M mutant bind to the GYF domain in a similar fashion. The terminal residues (Ser1, His2, His9 and Val11) are not involved in hydrophobic or electrostatic interactions that are crucial for

the binding in either NMR structure or conformations sampled in the simulations. Therefore we conclude that the tolerated mutations of these positions will not change the binding interface of the GYF domain significantly and the mutated peptides will bind to the domain analogously to the wt peptide.

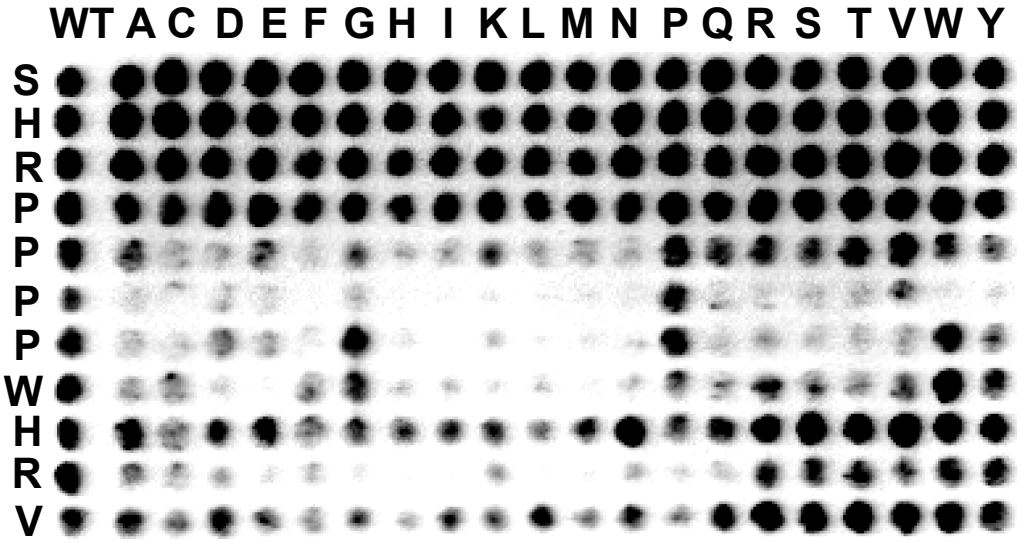


Figure 7: Substitution analysis of the SHRPPPPWHR peptide binding to the GYF domain. All single substitution analogues of the peptide were synthesized on a cellulose membrane. The single letter code above each column marks the amino acid that replaces the corresponding wild-type residue, while the row defines the position of the substitution within the peptide. Spots in the most left column (WT) have identical sequences and represent the wild type peptide. The membrane was incubated with a GST-GYF construct of CD2BP2. Bound protein was detected with an anti-GST primary antibody and a horse-radish peroxidase coupled secondary antibody. The relative spot intensities correlate qualitatively with the binding affinities (*45*)

G8X_DOCK - new binding modes. In the cluster analysis of the G8W_DOCK simulation, two large clusters were found: the first cluster covers the first 12 ns of the simulation while the second covers the simulation from 12 ns to 30 ns. Figure 6a shows the superposition of the representative conformations of each cluster. The most important differences between the two clusters are that the interacting prolines in the peptide are shifted one position: Pro5 and Pro6 now insert into the binding pocket instead of Pro6 and Pro7. All four prolines in the peptide are rotated clockwise when viewed from C-terminal to N-terminal. Interestingly, the orientations of the remaining residues were kept and show only a sight translation towards the C-terminus. When looking at the interactions between the peptides and the GYF domain, the electrostatic attraction between the positively charged residues Arg3 and Arg10 of the peptide and the negatively charged residues Glu31 and Glu9 of the domain were maintained in both clusters. The hydrogen bonds between the backbone oxygen of peptide Pro4 and H_ε of the domain Trp28 side-chain and between the backbone oxygen of peptide Pro7 and H_ε of the domain Trp8 side-chain were kept in the first cluster, while the acceptor atoms were shifted to peptide Arg3 and peptide Pro6 in the second cluster as a consequence of the translation. To obtain experimental backup for this proposed

binding mode, a peptide substitution analysis with the SHRPPPPWHR peptide was performed. In this experiment each amino acid of the peptide is individually exchanged against all other naturally occurring amino acids. Thereby the contribution of individual amino acids to the binding event can be estimated. Figure 7 shows the result of this experiment. Clearly, the importance of the PPW motif is suggested by the observed spot intensities, since mutations at these positions are mostly not compatible with detectable interactions. Furthermore, the second proline of the motif, which is exposed to solvent when bound in the original wild-type conformation, also shows considerable conservation. This is in agreement with the alternative binding mode suggested by the MD simulations. The second proline in the new binding mode would now contact the domain directly and thereby contribute to binding.

To further validate this new finding, we also performed a cluster analysis on the G8R_DOCK simulation. Supporting, the same motion was also found in this simulation and the main conformation occupying the entire 30 ns was similar to the shifted register binding observed in the second cluster of the G8W_DOCK simulation (see Figure 6b).

2.5.3 Implications of the alternative binding modes

To investigate the function of this alternative binding mode in depth and to test whether it only occurs for mutant peptides, we carried out five 20 ns MD simulations (ALT_GYF) of the wild-type complex starting from the same register shifted mode found in the second cluster of the G8W_DOCK simulation with different random seeds for the generation of initial atomic velocities. While the systems' overall behavior may be a bit different (the C_α RMSD with respect to the starting coordinates are stable around 2 Å in four simulations but reached 3 – 3.5 Å in the fifth simulation due to some structural changes at the C-terminus of the GYF domain), interestingly, the alternative binding mode of the peptide is well maintained in all five simulations of the wild-type complex. In cluster analyses carried out after the simulations only one dominant cluster was found in each simulation. An overlap of all the representative conformations is shown in Figure 6c. Although some lateral movements are observed between the individual conformations (in part due to different relative positions of the GYF domain used for superposition), note that Pro7 never shifts to the position of Pro6. Knowing that polyproline peptides bind to SH3 domains in both directions, we believe that the motion found in the G8W_DOCK and G8R_DOCK simulations is probably due to a transition between two alternative binding modes. This 'screw-like' rotation-translation motion (Figure 6d) or the transition between different binding modes can decrease the entropic penalty of the binding without affecting the specificity. Providing two alternative binding modes for a peptide should, theoretically, provide additional stability for the bound conformation due to the larger number of states accessible inside the minimum energy well of the bound state (Figure 8). Another possible function of this motion may be related to the binding mechanism: the peptides and the GYF domain first attract each other by the long-range electrostatic interactions between the charged residues, and then the peptides bind or leave the binding interface of GYF domain by this "screw like" motion along the interface. In addition, such screw-like motions may allow for a kinetically favorable binding process by "stripping off water molecules" upon binding and/or unbinding. Furthermore, these sites might act as delocalized anchors within

protein associations that rely on fast structural rearrangements within the context of eukaryotic signal transduction.

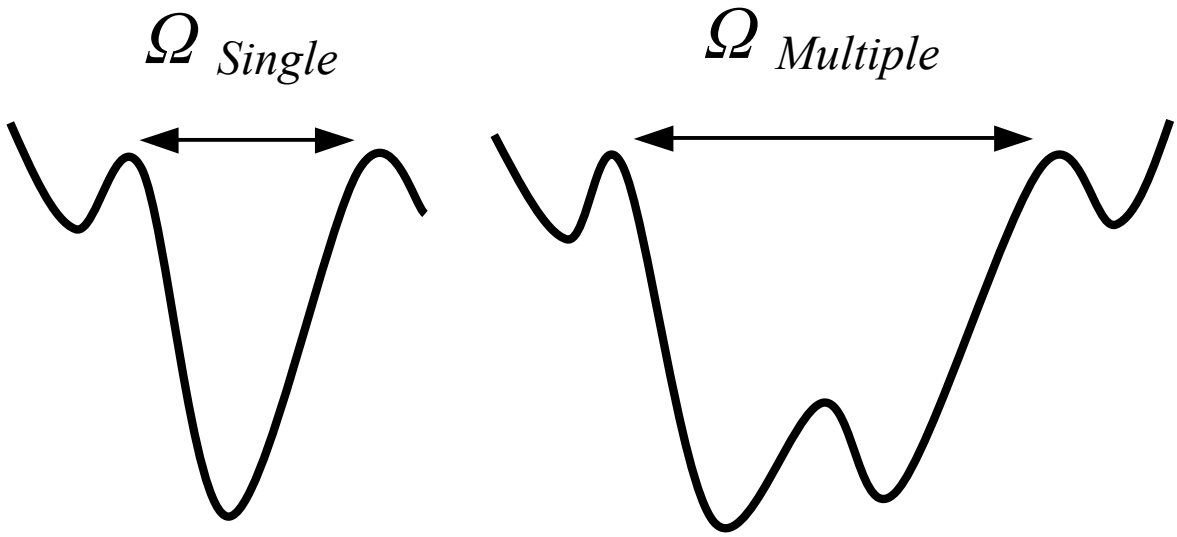


Figure 8: Funnel representation of the energy landscape for the binding of peptides to GYF domain. On the left is the single-mode binding; on the right is multiple-mode binding. In the multiple-mode binding, the system can access significantly more states than in the single-mode binding and thus becomes entropically more favourable.

2.5.4 Consistency between NMR experiments and theoretical calculations

Binding analysis of the GYF domain in regard to the mutant and wild-type peptides and chemical shift changes mapping (see Figure 4b) show the most significant difference between the wild-type complex and the G8W mutant to be the "backward" chemical shift changes of the backbone HN of Trp8. The chemical shifts of N-H group are mostly influenced by the local environment, e.g. alteration of H-bond strength, weaker effects of internal geometry, or by covalently linked aromatic groups. Therefore, the "backward" chemical shift change indicates that the local environment of Trp8 in the mutant complex is closer to the free form of the GYF domain. Theoretical methods of predicting chemical shifts are still not applicable for systems as used in the present study: for *ab initio* approaches, averaging of many conformations and including solvent effects for large systems is still too expensive, while for knowledge-based approaches the required specificity may not be obtained (0.1 p.p.m. of hydrogen) (*57-59*). Hence, direct comparison of the structure is more straightforward in this case. When looking at the structures of the free GYF domain (*10*), the wild-type complex (*14*) and the mutant complex (simulation), the backbone HN of Trp8 is exposed to the solvent (SASA > 0) in the free form of GYF domain, while it is buried in the two complexes (wild-type and mutant). However, in the G8W mutant complex, an internal hydrogen bond is formed between the HN of Trp8

(domain) and the carbonyl oxygen of Pro6/Pro7 (peptide); the distances between hydrogen and oxygen being 2.2 ± 0.3 Å. In the wild-type complex, no acceptor (oxygen atom) is found within 3.7 Å from the HN of Trp8 (domain). The newly formed hydrogen bond provides a similar local environment of the Trp8-HN of the GYF domain in the mutant complex as in the free form of the GYF domain. This is consistent with the "backward" chemical shift change of the HN of Trp8 of the GYF domain.

The average inter-molecular distances observed in the MD simulation of the wild-type complex started from the two different binding modes and both are in agreement with the distance restraints derived from the NOE data that were used for the calculation of the experimental structure. The stability of both conformations in MD simulations suggests that these alternate binding modes are possible for the wild-type peptide. They may be hard to distinguish experimentally since the NMR distance restraints comply with both conformations.

2.6 Conclusion

By using molecular modeling and MD simulations, totaling 450 nanoseconds of simulation time, we studied the solvent conformation of the wild-type and mutant polyproline peptides that bind to the GYF domain. We found that the peptides formed PPII helix conformations even in absence of the GYF domain. These results agree well with recent experimental and theoretical studies on polypeptides with or without prolines and indicate that the formation of a PPII helix of the peptide is not induced by the binding processes alone.

Based on the simulations of the wild-type and mutated complexes, and on our previous knowledge from NMR experimental studies of the GYF domain-ligand interaction, we modeled the general binding mode of polyproline peptides to the GYF domain. The hydrophobic interactions between the peptide residues Pro6 and Pro7, and binding pocket as well as the electrostatic attractions between the peptide residues Arg3 and Arg10, and the domain residues Glu31 and Glu9 play crucial roles in the binding.

Peptide docking and subsequent MD simulations of the G8X mutants identified an alternative binding mode, where a shift in register for the interacting prolines was observed. These results agree qualitatively well with NMR chemical shift mapping experiments and indicate dynamic processes to be important for proline-rich sequence recognition. Possibly, such gliding motions along long proline-rich sequences decrease the entropic penalty of binding while still keeping a certain degree of specificity.

References

1. Kay, B. K., Williamson, M. P., and Sudol, P. (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains, *Faseb J. 14*, 231-241.
2. Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S. B., Gong, F. C., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J. Y., Li, Z. Y., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. Q. H., Zhong, F., Zhong, W. Y., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. (2000) Comparative genomics of the eukaryotes, *Science 287*, 2204-2215.
3. Zarrinpar, A., Bhattacharyya, R. P., and Lim, W. A. (2003) The structure and function of proline recognition domains, *Sci. STKE. RE8*.
4. Carlsson, L., Nystrom, L. E., Sundkvist, I., Markey, F., and Lindberg, U. (1977) Actin polymerizability is influenced by profilin, a low molecular weight protein in non-muscle cells, *J. Mol. Biol. 115*, 465-483.
5. Mayer, B. J., Hamaguchi, M., and Hanafusa, H. (1988) A novel viral oncogene with structural similarity to phospholipase C, *Nature 332*, 272-275.
6. Stahl, M. L., Ferenz, C. R., Kelleher, K. L., Kriz, R. W., and Knopf, J. L. (1988) Sequence similarity of phospholipase C with the non-catalytic region of src, *Nature 332*, 269-272.
7. Bork, P., and Sudol, M. (1994) The Ww Domain - a Signaling Site in Dystrophin, *Trends Biochem.Sci. 19*, 531-533.
8. Niebuhr, K., Ebel, F., Frank, R., Reinhard, M., Domann, E., Carl, U. D., Walter, U., Gertler, F. B., Wehland, J., and Chakraborty, T. (1997) Novel proline-rich motif present in ActA of Listeria monocytogenes and cytoskeletal proteins is the ligand for the EVH1 domain, a protein module present in the Ena/VASP family, *Embo J. 16*, 5433-5444.
9. Nishizawa, K., Freund, C., Li, J., Wagner, G., and Reinherz, E. L. (1998) Identification of a proline-binding motif regulating CD2- triggered T lymphocyte activation, *Proc. Natl. Acad. Sci. U. S. A. 95*, 14897-14902.
10. Freund, C., Dotsch, V., Nishizawa, K., Reinherz, E. L., and Wagner, G. (1999) The GYF domain is a novel structural fold that is involved in lymphoid signaling through proline-rich sequences, *Nat. Struct. Biol. 6*, 656-660.
11. Sancho, E., Vila, M. R., Sanchez-Pulido, L., Lozano, J. J., Paciucci, R., Nadal, M., Fox, M., Harvey, C., Bercovich, B., Loukili, N., Ciechanover, A., Lin, S. L., Sanz, F., Estivill, X., Valencia, A., and Thomson, T. M. (1998) Role of UEV-1, an inactive variant of the E2 ubiquitin- conjugating enzymes, in in vitro differentiation and cell cycle behavior of HT-29-M6 intestinal mucosecretory cells, *Mol. Cell. Biol. 18*, 576-589.
12. Pornillos, O., Alam, S. L., Davis, D. R., and Sundquist, W. I. (2002) Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein, *Nat. Struct. Biol. 9*, 812-817.
13. Myllyharju, J., and Kivirikko, K. I. (1999) Identification of a novel proline-rich peptide-binding domain in prolyl 4-hydroxylase, *Embo J. 18*, 306-312.
14. Freund, C., Kuhne, R., Yang, H. L., Park, S., Reinherz, E. L., and Wagner, G. (2002) Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules, *Embo J. 21*, 5985-5995.
15. Freund, C., Kuhne, R., Park, S., Thiemke, K., Reinherz, E. L., and Wagner, G. (2003) Structural investigations of a GYF domain covalently linked to a proline-rich peptide, *J. Biomol. NMR 27*, 143-149.
16. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: Re-assessing the

protein structure-function paradigm, *J. Mol. Biol. 293*, 321-331.
17. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity - linking function and disorder, *Nat. Biotechnol. 19*, 805-806.
18. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. R., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C. H., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, M., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J. Mol. Graph. 19*, 26-59.
19. Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T., and Rose, P. W. (2003) Simulating disorder-order transitions in molecular recognition of unstructured proteins: Where folding meets binding, *Proc. Natl. Acad. Sci. U. S. A. 100*, 5148-5153.
20. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54-60.
21. Hamburger, J. B., Ferreon, J. C., Whitten, S. T., and Hilser, V. J. (2004) Thermodynamic Mechanism and Consequences of the Polyproline II (P(II)) Structural Bias in the Denatured States of Proteins, *Biochemistry 43*, 9790-9799.
22. Blanco, F. J., Rivas, G., and Serrano, L. (1994) A Short Linear Peptide That Folds into a Native Stable Beta- Hairpin in Aqueous-Solution, *Nat. Struct. Biol. 1*, 584-590.
23. Vila, J. A., Baldoni, H. A., Ripoll, D. R., Ghosh, A., and Scheraga, H. A. (2004) Polyproline II helix conformation in a proline-rich environment: A theoretical study, *Biophys. J. 86*, 731-742.
24. Rucker, A. L., Pager, C. T., Campbell, M. N., Qualls, J. E., and Creamer, T. P. (2003) Host-guest scale of left-handed polyproline II helix formation, *Proteins 53*, 68-75.
25. Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F., and Creamer, T. P. (2001) Host-guest study of left-handed polyproline II helix formation, *Biochemistry 40*, 14376-14383.
26. Williamson, M. P. (1994) The Structure and Function of Proline-Rich Regions in Proteins, *Biochem. J. 297*, 249-260.
27. Shi, Z. S., Olson, C. A., Rose, G. D., Baldwin, R. L., and Kallenbach, N. R. (2002) Polyproline II structure in a sequence of seven alanine residues, *Proc. Natl. Acad. Sci. U. S. A. 99*, 9190-9195.
28. Rucker, A. L., and Creamer, T. P. (2002) Polyproline II helical structure in protein unfolded states: Lysine peptides revisited, *Protein Sci. 11*, 980-985.
29. Woody, R. W. (1992) Circular dichroism and conformation of unordered polypeptides, *Adv. Biophys. Chem. 2*, 37-79.
30. Tiffany, M. L., and Krimm, S. (1972) Effect of temperature on the circular dichroism spectra of polypeptides in the extended state, *Biopolymers 11*, 2309-2316.
31. Tiffany, M. L., and Krimm, S. (1968) Circular dichroism of poly-L-proline in an unordered conformation, *Biopolymers 6*, 1767-1770.
32. Asher, S. A., Mikhonin, A. V., and Bykov, S. (2004) UV Raman demonstrates that alpha-helical polyalanine peptides melt to polyproline II conformations, *J. Am. Chem. Soc. 126*, 8433-8440.
33. Kentsis, A., Mezei, M., Gindin, T., and Osman, R. (2004) Unfolded state of polyalanine is a segmented polyproline II helix, *Proteins 55*, 493-501.
34. Mezei, M., Fleming, P. J., Srinivasan, R., and Rose, G. D. (2004) Polyproline II helix is the preferred conformation for unfolded polyalanine in water, *Proteins 55*, 502-507.
35. Chellgren, B. W., and Creamer, T. P. (2004) Short sequences of non-proline residues can adopt the polyproline II helical conformation, *Biochemistry 43*, 5864-5869.
36. Creamer, T. P. (1998) Left-handed polyproline II helix formation is (very) locally driven, *Proteins 33*, 218-226.
37. Sreerama, N., and Woody, R. W. (1999) Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(Pro)II conformations, *Proteins 36*, 400-406.
38. Pappu, R. V., and Rose, G. D. (2002) A simple model for polyproline II structure in unfolded states of alanine-based peptides, *Protein Sci. 11*, 2437-2455.
39. Stapley, B. J., and Creamer, T. P. (1999) A survey of left-handed polyproline II helices, *Protein Sci. 8*, 587-595.
40. Kofler, M., Heuer, K., Zech, T., and Freund, C. (2004) Recognition sequences for the GYF domain reveal a possible spliceosomal function of CD2BP2, *J. Biol. Chem. 279*, 28292-28297.
41. Feng, S. B., Chen, J. K., Yu, H. T., Simon, J. A., and Schreiber, S. L. (1994) 2 Binding

Orientations for Peptides to the Src Sh3 Domain - Development of a General-Model for Sh3-Ligand Interactions, *Science 266*, 1241-1247.
42. Goddard, T. D., and Kneller, D. G., SPARKY 3, University of California, San Francisco.
43. Frank, R. (1992) Spot-synthesis – an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support, *Tetrahedron 48,* 9217-9232.
44. Kramer, A. and Schneider-Mergener J. (1998) Synthesis and screening of peptide libraries on continuous cellulose membrane supports, *Methods Mol. Biol. 87,* 25-39
45. Kramer, A., Reineke, U., Dong, L., Hoffmann, B., Hoffmuller, U., Winkler, D., Volkmer-Engert, R., and Schneider-Mergener, J. (1999) Spot synthesis: observations and optimizations, *J. Pept. Res. 54,* 319-327).
46. Heuer, K., Kofler, M. Langdon, M., Thiemke, K. and Freund, C. (2004) Structure of a helically extended SH3 domain of the T cell adapter protein ADAP. *Structure 12*, 603-610.
47. van der Spoel, D., van Drunen, R., and Berendsen, H. J. C., Groningen Machine for chemical simulation. (Department of Biophysical Chemistry, BIOSON Research Institute Nijenborgh 4 NL-9717 AG Groningen, 1994).
48. Jorgensen, W. L., Maxwell, D. S., and TiradoRives, J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc. 118*, 11225-11236.
49. Ren, P. Y., and Ponder, J. W. (2003) Polarizable atomic multipole water model for molecular mechanics simulation, *J. Phys. Chem. B 107*, 5933-5947.
50. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys. 79*, 926-935.
51. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. (1997) LINCS: A linear constraint solver for molecular simulations, *J. Comput. Chem. 18*, 1463-1472.
52. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A Smooth Particle Mesh Ewald Method, *J. Chem. Phys. 103*, 8577-8593.
53. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 684-3690.
54. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol. 261*, 470-489.
55. Cowan, P. M., McGavin, S., and North, A. C. (1955) The polypeptide chain configuration of collagen, *Nature 176*, 1062-1064.
56. Parker, D., Rivera, M., Zor, T., Henrion-Caude, A., Radhakrishnan, I., Kumar, A., Shapiro, L. H., Wright, P. E., Montminy, M., and Brindle, P. K. (1999) Role of secondary structure in discrimination between constitutive and inducible activators, *Mol. Cell. Biol. 19*, 5601-5607.
57. Neal, S., Nip, A. M., Zhang, H. Y., and Wishart, D. S. (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts, *J. Biomol. NMR 26*, 215-240.
58. Meiler, J. (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks, *J. Biomol. NMR 26*, 25-37.
59. Gronwald, W., Boyko, R. F., Sonnichsen, F. D., Wishart, D. S., and Sykes, B. D. (1997) ORB, a homology-based program for the prediction of protein NMR chemical shifts, *J. Biomol. NMR 10*, 165-179.
60. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics, *J. Mol. Graph. 14*, 33-38.

Chapter 3

Cyclophilin A Binds to Linear Peptide Motifs Containing a Consensus That Is Present in Many Human Proteins*

(published in *J. Biol. Chem.*, **280**, 23668-23674 (2005))

3.1 Summary

Cyclophilin A (CypA) is a peptidyl-prolyl *cis*/*trans*- isomerase that is involved in multiple signaling events of eukaryotic cells. It might either act as a catalyst for prolyl bond isomerization, or it can form stoichiometric complexes with target proteins. We have investigated the linear sequence recognition code for CypA by phage display and found the consensus motif FGPXLp to be selected after five rounds of panning. The peptide FGP- DLPAGD showed inhibition of the isomerase reaction and NMR chemical shift mapping experiments highlight the CypA interaction epitope. Ligand docking suggests that the peptide was able to bind to CypA in the *cis*- and *trans*-conformation. Protein Data Bank searches reveal that many human proteins contain the consensus motif, and several of these protein motifs are shown to interact with CypA in vitro. These sequences represent putative target sites for binding of CypA to intracellular proteins.

3.2 Introduction

Cyclophilin A (CypA) is a ubiquitously expressed protein that has been found in a variety of functional contexts. On the one hand, CypA serves as immunophilin and binding partner for the immunosuppressive drug cyclosporin A (*1*), and on the other hand CypA was found to catalyze the *cis*/*trans*- isomerization of proline imide bonds in peptides (*2*). Peptidyl-prolyl *cis/trans*- activity of CypA was shown to be relevant for protein folding in vitro (*3*), but it has been difficult to prove the relevance of catalysis in vivo. Recent experiments suggest that the tyrosine kinase Itk is regulated by the catalytic activity of CypA (*4*) and for the human immunodeficiency virus

* This work is in collaboration with the group of Dr. Christian Freund in the Freie Universität Berlin. In this chapter, only the computational part and one figure from the experimental part (for comparison reason) of this work are presented.

(HIV) Vpr protein *cis/trans*- interconversion of a critical proline residue might be catalyzed by CypA (*5*). The role of CypA as stoichiometric binding partner has also gained recent interest in the light of HIV infectivity and T cell signal- ing. For certain HIV-1 strains, the ability of the capsid protein (CA) to bind CypA correlates well with the infectivity of these strains (*6*). Structural analysis shows that the GP-dipeptide of the CA86–93 loop is deeply inserted into the CypA binding site, and mutation of glycine to alanine reduces ground state binding, presumably fostering the *cis/trans*-interconversion rate (*7*). Therefore, binding of the X-P-dipeptide bond may result in stable complex formation or transient interaction and cataly- sis, depending on the sequence or conformational context of the critical proline. The sequence context for catalysis by CypA has been investigated in detail (*8*) and revealed no stringent re-quirements for the nature of the amino acid flanking the cen- tral proline. For immunophilins as binding modules it was suggested that peptide conformation is central to the formation of complexes (*9*); however, a global analysis of the sequences binding to CypA is missing. Here we apply phage display from a randomized 9-mer peptide library to map the recognition code for linear peptide sequences that interact with CypA. We iden- tify the consensus sequence as FGPXLp and confirm the importance of the individual amino acids for the peptide FGPDLPAGD. The latter peptide is an active site inhibitor, and NMR spectroscopy shows that the binding epitope overlaps with the interaction site of cyclosporin A and other CypA ligands. The recognition signature is present in a number of human proteins, and peptides derived from these proteins bind to CypA when spotted onto a nitrocellulose membrane. The identification of novel CypA binding sites sets the stage for the detection of yet unknown CypA interaction partners. As a first example, we show that a phage display-derived peptide motif is bound by CypA in the context of the entire cytoplasmic domain of the T cell adhesion molecule CD2.

3.3 Materials and Methods

Modeling of the CypA-Peptide Complex—Since the cyclophilin-bound CAN loop has a similar sequence (XGPX) as the peptide investigated here, we used the x-ray structure of the complex CypA/HIV-1 CAN, Protein Data Bank entry 1M9C (*7*), as a template for modeling. The FlexX package (*10*) was used for docking of the peptides FGPDLPAGD and FGPDLP to CypA. All atoms of CypA and Pro90 (chain D) were fixed during the docking, and the remaining residues of the peptide were subjected to exhaustive conformational analysis. Two complex conformations displaying the highest binding energy score were analyzed in more detail. After docking, the two candidate complexes were solvated in cubic boxes using TIP3P water molecules (*11*) with an initial minimum distance of at least 8 Å between the boundaries of the box and the nearest solute atom. The systems were first optimized by 500 steps energy minimization each using the GROMACS3.14 package (*12*) and the OPLSAA force field (*13*). Subsequently, a 1-ns molecular dynamics simulation was carried out for each candidate with restrained positions for all the template atoms in CypA and Pro90 (chain D). Finally, two complete 10-ns molecular dynamics simulations with no positional restraint were performed to fully optimize the structure of the two candidates. The LINCS procedure (*14*) was applied to constrain all bond lengths. The time step of the simulation was set to 2 fs. A 9 Å cutoff was used for the short-range

non-bonded interactions and the lists of nonbonded pairs were updated every 10 steps. The Particle Mesh Ewald method (*15*) with a grid size of 1.2 Å was used to calculate long-range electrostatic interactions. In the simulations, temperature and pressure were maintained by weak coupling to an external bath (*16*). All simulations were performed at 300 K temperature. Cluster analysis was applied after the simulations. Main-chain atoms and Catoms were selected for calculating the root mean square deviation matrix. The root mean square deviation cutoff was set to 1.0 Å.

3.4 Results

3.4.1 Substitution analysis of the peptide FGPDLPAGD

For comparison of the modeling results, Figure 1 shows the results from the substitution analysis of the peptide FGPDLPAGD. All possible single site substitution variants of the peptide were synthesized on a nitrocellulose membrane. The single letter code above each column indicates the amino acid that replaces the corresponding wild-type residue; the row defines the position of the substitution within the peptide. Spots in the first column display wild type peptide in all cases. The membrane was incubated with GST-CypA, and bound protein was detected with an anti-GST primary antibody and a horseradish peroxidase coupled secondary antibody. The relative spot intensities correlate qualitatively with the binding affinities.(*17*)

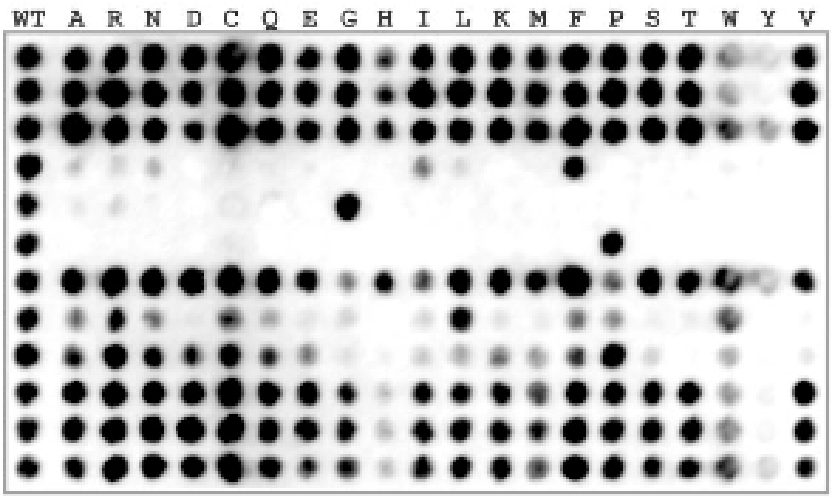


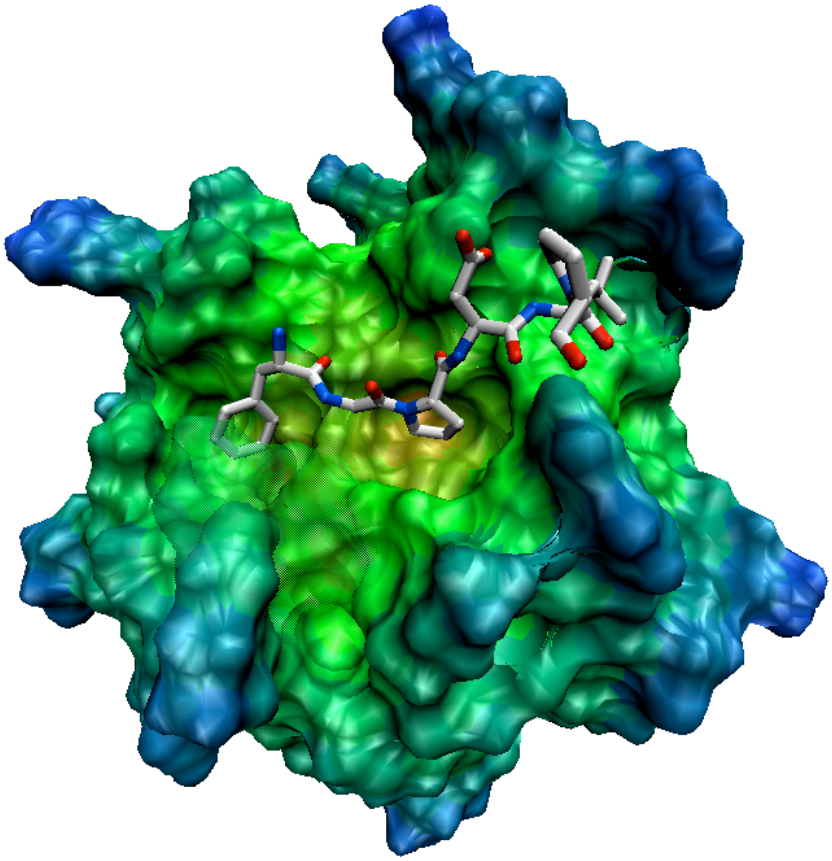
Figure 1: substitution analysis of the peptide FGPDLPAGD.

3.4.2 Model of CypA bound to the phage display-derived peptide

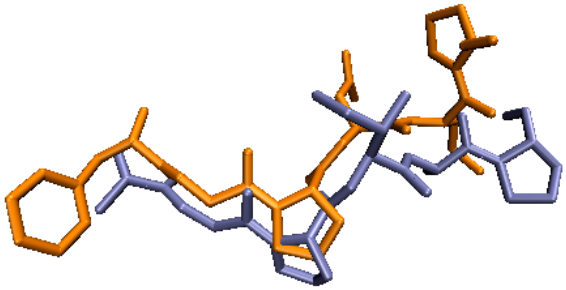
The docked and optimized model of CypA bound to the FGPDLP motif of the phage display-derived peptide as well as the superposition of the *cis*- and *trans*-variants of the peptide are shown in Fig. 3, A and C. To analyze possible interactions of the ligand with CypA a cutoff was set to 3.4 and 5.0 Å for hydrogen bonds and van der Waals contacts, correspondingly. The GP motif adopts a trans-conformation and fits

well to the hydrophobic pocket defined by residues Ile57, Phe60, Met61, Ala101, Ala103, and Leu122, involving also Gln63, Asn102, Phe113, and His126 (Figure 2a). The glycine at position *i*–1 is essential, since any other residues would result in a clash between the ligand side-chain and CypA. The phenylalanine at position *i*–2 fits well into the indentation formed by Lys82, Ala101, Asn102, Ala103, Thr107, Asn108, Gly109, and Gln111. These results are in agreement with the peptide substitution analysis, which shows the FGP segment to be exclusively required for high affinity binding. At position *i*–1, the side chain of aspartic acid points toward the solvent, which is in accordance with the mutational analysis (Figure 1), where most amino acid substitutions are tolerated at this position. The leucine residue at position *i*+2 is involved in van der Waals interactions with Ile57, Asn71, and Arg148 of CypA; however, membrane spot analysis suggests that other hydrophobic amino acids at this position can support similar interactions (Figure 1). Superposition of the modeled peptide with the x-ray structure shows a good fit of the first five residues (FGPDL) to the CA fragment despite the difference in the amino acid sequence (Figure 2b). The proline at position *i*+3 in our model points toward the solvent, while it is oriented toward CypA in the experimental structure (Figure 2b). Figure 3c shows the superposition of the two docking models with trans- and *cis*-conformations of the GP motif. It can be seen that the CypA binding site could accommodate both variants, since the major conformational change affects G_{i-1}, whereas the hydrophobic interactions can be maintained. Finally NMR studies of the peptide in complex with CypA will allow to experimentally validating this model.

(a)



(b)



(c)

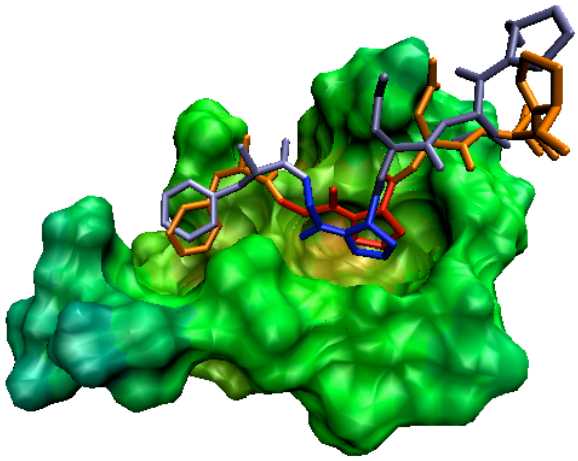


Figure 2: Modeling of the CypA-bound peptide. (a) Modeled complex of CypA with the peptide segment FGPDLP. Shown is the final complex obtained from docking, subsequent refinement by 10 ns of molecular dynamics simulation, and clustering analysis. The surface of CypA is colored according to solvent accessibility: buried (*orange*) and exposed (*deep blue*); the peptide is shown as sticks in Corey-Pauling-Koltun colors. (b) Superposition of the modeled peptide (*orange*) and the segment Ala88–Pro93 of the CA^N-ligand, Protein Data Bank entry 1M9C (*indigo*). (c) Superposition of the modeled *trans*- and *cis*-conformers of the ligand in the binding pocket of CypA. Colors of the surface are as described for (a). The GP motif is shown in *red* (*trans*) and *blue* (*cis*), and the rest of the ligand is *orange* (*trans*) or *indigo* (*cis*).

3.5Conclusion

The substitution analysis (phage display) identified the linear sequence recognition code for CypA and found the consensus motif FGPXLp. The modeled complex structure agrees very well with the results form phage display experiments and gives an explanation of the specific binding motif from structural and interaction points of view.

References

1. Handschumacher, R. E., Harding, M. W., Rice, J., Drugge, R. J., and Speicher, D. W. (1984) Cyclophilin: a specific cytosolic binding protein for cyclosporin A, *Science* 226, 544 –547
2. Fischer, G., Bang, H., and Mech, C. (1984) Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides, *Biomed. Biochim. Acta* 43, 1101–1111
3. Lang, K., Schmid, F. X., and Fischer, G. (1987) Catalysis of protein folding by prolyl isomerase, *Nature* 329, 268 –270
4. Brazin, K. N., Mallis, R. J., Fulton, D. B., and Andreotti, A. H. (2002) Regulation of the tyrosine kinase Itk by the peptidyl-prolyl isomerase cyclophilin A, *Proc. Natl. Acad. Sci. U. S. A.* 99, 1899 –1904
5. Bruns, K., Fossen, T., Wray, V., Henklein, P., Tessmer, U., and Schubert, U. (2003) Structural Characterization of the HIV-1 Vpr N Terminus EVIDENCE OF cis/trans-PROLINE ISOMERISM, *J. Biol. Chem.* 278, 43188 – 43201
6. Wiegers, K., and Krausslich, H. G. (2002) Differential Dependence of the Infectivity of HIV-1 Group O Isolates on the Cellular Protein Cyclophilin A, *Virology* 294, 289 –295
7. Howard, B. R., Vajdos, F. F., Li, S., Sundquist, W. I., and Hill, C. P. (2003) Structural insights into the catalytic mechanism of cyclophilin A, *Nat. Struct. Biol.* 10, 475– 481
8. Harrison, R. K., and Stein, R. L. (1992) Mechanistic studies of enzymic and nonenzymic prolyl cis-trans isomerization, *J. Am. Chem. Soc.* 114, 3464 –3471
9. Ivery, M. T. G. (2000) Immunophilins: Switched on protein binding domains?, Med. Res. Rev. 20, 452– 484
10. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.* 261, 470 – 489
11. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.* 79, 926 –935
12. Lindahl, E., Hess, B., and van der Spoel, D. (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis, *J. Mol. Model.* 7, 306 –317
13. Jorgensen, W. L., and Tiradorives, J. (1988) The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc.* 110, 1657–1666
14. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations, *J. Comp. Chem.* 18, 1463–1472
15. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald potential, *J. Chem. Phys.* 103, 8577– 8593
16. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* 81, 3684 –3690
17. Kramer, A., Reineke, U., Dong, L., Hoffmann, B., Hoffmuller, U., Winkler, D., Volkmer-Engert, R., and Schneider-Mergener, J. (1999) Spot synthesis: observations and optimizations *J. Pept. Res.* 54, 319 –327

Chapter 4

Dynamical Binding of Proline-rich Peptides to their Recognition Domains

(published in *Biochim. Biophys. Acta - Proteins and Proteomics*, **1754**, 232-238 (2005))

4.1 Summary

Recognition of proline-rich sequences plays an important role for the assembly of multi-protein complexes during the course of eukaryotic signal transduction and is mediated by a set of protein folds that share characteristic features. For many complex systems containing proline-rich sequences, multiple binding modes have been found by theoretical and/or experimental studies. In this review, we discuss the different binding modes as well as the correlated dynamics of the peptides and their recognition domains, and some implications to their biological functions. Furthermore, we give an outlook of the systems in the field of systems biology.

4.2 Introduction

Intracellular protein domains recognizing proline-rich sequences (PRS) play a pivotal role in biological processes that require the coordinated assembly of multi-protein complexes(*1*). One example are Src kinases, where the terminal SH3 domain recognizes a long proline-rich stretch linking the nearby SH2 domain with the catalytic kinase domain(*2*). In vertebrate genomes, PRS are predicted to be among the most abundantly expressed amino acid sequence motifs (*3*) and this corresponds to an increasing number of proteins that acquired PRS-recognition domains during the course of evolution (*4*). Up to now, the super-family of proline-rich sequence recognition domains consists of profilin (*5*), the SH3 (*6, 7*), the WW(*8*), the EVH1 (*9*), the GYF(*10, 11*), the UEV(*12, 13*) and probably the ligand binding domain of prolyl-4-hydroxylase (*14*). For each of these domains a set of conserved aromatic amino acid residues is important for peptide binding (see Figure 1).

The PRS and their recognition domains as well as the common structure function relationships have been recently reviewed several times(*4, 15-17*). Here, we will focus on the dynamics and conformational variability for both interaction partners, the

roles of these changes in the binding process as well as their potential biological advantages. At the end of this mini-review, we take a look into the future and point out fruitful areas, e.g. in systems biology, for further studies.

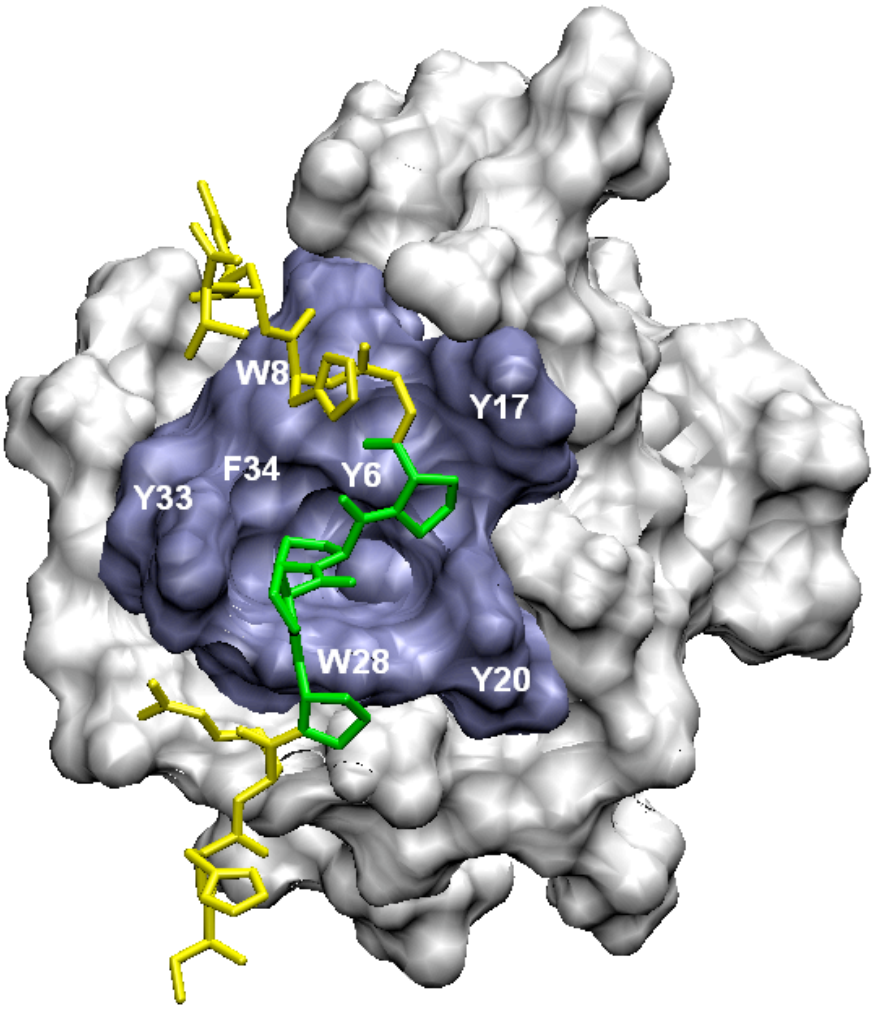


Figure 1: NMR structure of GYF domain with wild-type peptide. The GYF domain is represented by its molecular surface; the peptide atoms are drawn as sticks. Residues forming the binding pocket are coloured in dark grey and labelled by their one-letter codes and sequence numbers. The four proline residues are coloured in green.

4.3 Proline and Proline-rich Sequences

Among the 20 naturally occurring amino acids, proline is the only one in which the side chain atoms form a pyrrolidine ring with the backbone atoms (see Figure 2a). This cyclic structure leads to some distinguished properties of proline: it induces conformational constraints among the atoms in the pyrrolidine ring, and it is the reason for the slow isomerization between *cis*/*trans* conformations (*18*) and for the

secondary structure preferences of proline-rich sequences (see below). Remarkably, the *cis*/*trans* preference of proline-X (where X is any amino acid) peptide bonds are different in different solvent environments(*19-23*).

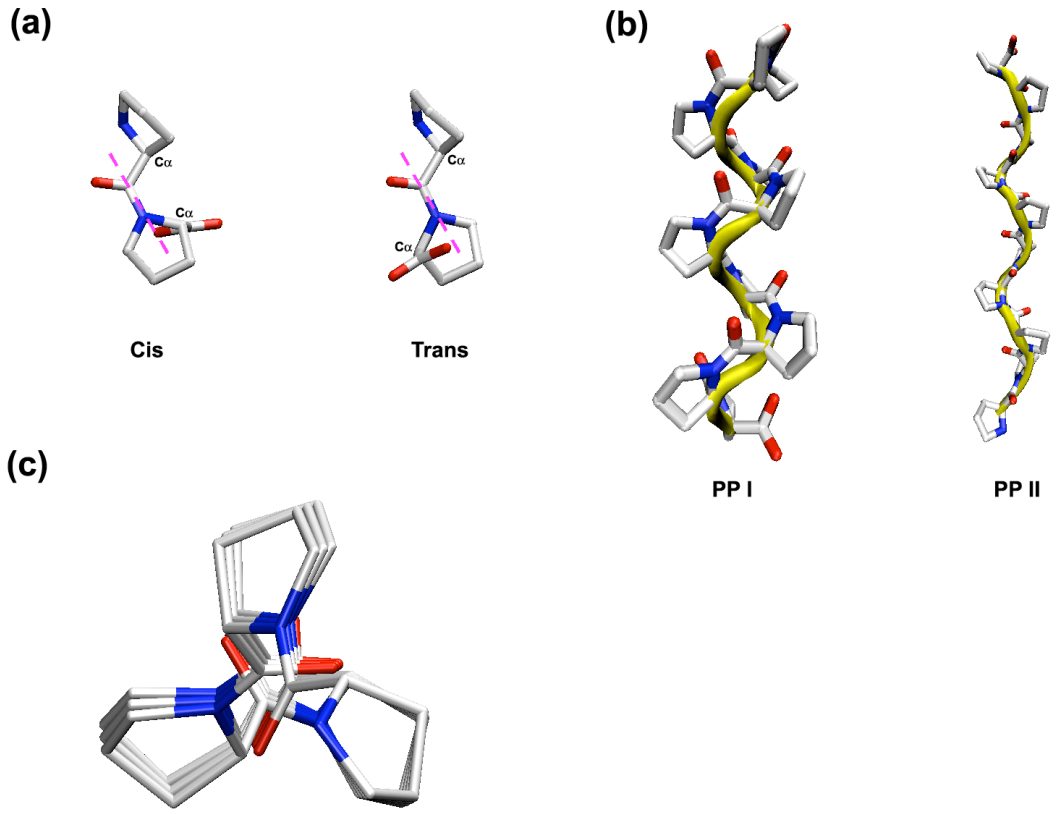


Figure 2: Structure of (a) the *cis* and *trans* proline residues; (b) the PPI and PPII helices; (c) the PPII helix viewed along the helical axis. Molecules are shown as sticks in Corey-Pauling-Koltun colors.

Type of helix	Phi (degree)	Psi (degree)	Omega (degree)	Num. residues per turn	Helical rise per residue (Å)	Helical pitch (Å/turn).
PPI[1]	-75	160	0	3.3	1.7	5.6
PPII[2]	-75	145	180	3.0	3.1	9.3

Table 1 Geometric properties of polyproline helices I and II. PPI helix is right-handed and PPII helix is left-handed.

Due to the special properties of the proline residue, the proline-rich sequences tend to form either of the two different secondary structures: PPI helices (in which all prolines are *cis* isomers) and PPII helices (where all prolines are *trans* isomers)(see Figure 2b and Table 1). The PPII helix is a left-handed helix with three residues per turn (see Figure 2b and 2c). It has a three-fold symmetry when viewed along the helical axis and every forth residue is in the same position (at a distance of 9.3 Å from each other). Along the same axis, the PPII helix also has a two-fold rotational pseudo symmetry(*4*). The side chains and backbone carbonyl groups are located in similar positions in both orientations along the backbone axis. This leads to the special property that e.g. SH3 domains may bind their PRS ligands in two orientations(*24*). Due to the lack of intramolecular hydrogen bonds, PPII helices are more flexible than a-helices(*25*) and the backbone groups are more accessible to the solvents. This also means that the PPII motifs are mostly located on the surface of proteins(*26*). All these geometric features are important when PPII helices bind to the recognition domains. Switching between *cis* and *trans* forms occurs either spontaneously and slow(*18*) or is catalyzed by *cis/trans* isomerases as cyclophilins (Cyp), FK506-binding proteins (FKBPs), and the parvulins(*27, 28*).

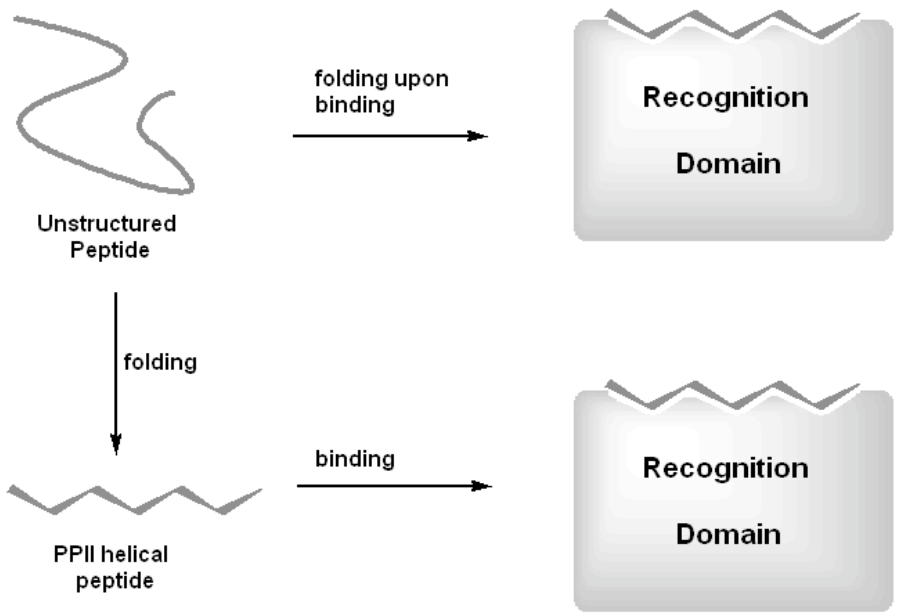
Rucker and Creamer argued that PPII is an energetically favorable option for oligopeptides because all backbone polar groups are well-solvated in this conformation in water, thus compensating for the lack of intramolecular hydrogen bonds.(*29*) Theoretical studies argued that PPII helices disrupt water organization less than b-sheet and a-helices, which makes them entropically favored as well.(*30, 31*) However, the dynamical features of PRS were emphasized by Scheraga and coworkers who claimed that the view that the optimal conformation of a polyproline-rich peptide is an ideal or canonical PPII helix in water is an oversimplification, and one should consider *cis-trans* isomerization of the proline peptide groups(*23*). In contrast to the aqueous environment, where either PPI or PPII seems possible, all PRS peptides known so far adopt a pure polyproline type II helix upon binding to the recognition domains. (*6, 7, 10, 11, 32*). For peptides only containing few prolines, the cis isomer is favorable as well (33-35) and the cis-trans isomerization in these scenarios is often connected to the functions of the proteins (*33, 35*). An impressive example for such conformational control is a proline-driven conformational switch within the Itk SH2 domain(*35*). Two structures of Itk SH2 determined by NMR spectroscopy corresponding to the cis and trans imide bond-containing conformers indicate that the heterogeneous Pro residue acts as a hinge modulating ligand recognition by controlling the relative orientation of protein-binding surfaces. Therefore, *cis-trans* isomerization of a single prolyl imide bond within the SH2 domain mediates conformer-specific ligand recognition. This plays a functional role in mediating distinct intermolecular interactions with exogenous signaling partners, e.g. cyclophilin A (CypA), and further influencing the T cell activation(*35, 36*).

4.4 Preformation of the PPII Helix for Unbound PRS

It has been recently recognized that many proteins contain long disordered segments in their functional states under physiological conditions(*37-41*). E.g. most of the polypeptide hormones are conformationally disordered in aqueous solution and fold

upon binding to their receptors(*40*). Such unstructured segments within large proteins provide ideal scaffolds for the interaction with several different targets and thereby help to assemble multi-protein complexes(*37-41*). For those proteins with unstructured segments, the coupling of binding and folding is expected favorable in terms of the binding free energy: the entropic penalty associated with the folding transition is counterbalanced by a large enthalpy of binding(*37, 42*). In those cases, the folding upon binding acts as a fine controller of the thermodynamic balance.

On the other hand, it has been shown by many experimental and theoretical studies that certain peptides, including proline-rich sequences, adopt preferred conformations in solution(*1, 23, 29, 43-51*). Therefore, it is a matter of ongoing discussion whether the PPII helix is such a preferred conformation for particular peptide sequences(*13, 29-31, 43, 46, 48, 49, 52-57*). A mechanistic description of the binding event has to distinguish whether the PPII helix conformation is preformed in the unbound peptides and binding to the recognition domains takes place in a "lock and key" mode or whether folding and binding occur in parallel, corresponding to an 'induced fit' model (see Scheme 1).



Scheme 1

We have previously studied the binding of wild type and some mutated PRS binding to the GYF adaptor domain by a combined theoretical (molecular dynamics simulations) and experimental (NMR and phage display) approach(*58*). The polyproline peptides considered in this study were found to be already folded into a PPII helix conformation in the unbound state and bind constitutively to the GYF domain. Obviously, this binding scenario is entropically more favorable than binding of unstructured peptides. The rigid PPII helix conformation of the unbound peptides studied is apparently intrinsically stable in solution and is also favorable for its specific binding motif. An experimental study addressed the binding of the

polyproline Sos peptide to the Sem-5 SH3 domain(*43*). They found that the PPII bias of unstructured peptides is driven by a favorable and significant enthalpy (DH) of –1.7 kcal mol-1 residue-1, which is partially offset by an unfavorable entropy (TDS) of –0.7 kcal mol-1 residue-1, relative to the ensemble of disordered conformations of the molecule. A similar example is the c-Myb oncoprotein, which adopts an α-helical conformation both complexed and uncomplexed with its target protein(*42*). Remarkably, binding of c-Myb to its target (residue 586-672 of CREB binding protein) is entropically favored (DS = +7.5 cal mol-1 K-1) while its favorable enthalpy change is small (DH = -4.1 kcal mol-1 K-1)(*37, 42*).

In conclusion, it appears that the conformation of unbound peptides may be fine-tuned for a particular functional range of peptide binding. On the one hand, Dyson and Wright proposed that unstructured proteins provide a large flexibility of binding reactions because they may adopt various structures upon binding to different partners(*40*). On the other hand, as examplified here for the GYF domain-ligand pair, the preformation of a peptide conformation might be well suited to guarantee the rapid formation of specific peptide-protein complexes within the dynamic settings of signal transduction.

4.5 Different Binding Modes and Their Roles for Binding and Function

A classic NMR study(*24*) showed that SH3 domains can bind proline rich ligands in two orientations, due to the twofold rotational pseudo symmetry of the PPII helix along the helical axis (as discussed in the first part of this review). Is such a scenario possible for other domains as well and what is the biological advantage of these different binding modes? Newer crystallographic evidence then showed that profilin, like SH3 domains, can bind proline-rich peptides in two distinct amide backbone orientations(*59*). As has been previously proposed for SH3-related functions, the ability of profilin to bind ligands in multiple orientations may control the organization of multi component signaling complexes, and provides a mechanism for the regulation of actin cytoskeleton assembly.

How far does this conformational flexibility of recognition extend? In some cases, as for the Itk SH2 domain, a single proline flip may result in an on/off control of binding events. On the other hand, Piotukh et al studied linear peptide motifs binding to CypA using both experimental (phage display, NMR) and theoretical (docking and molecular dynamics simulation) approaches. They predicted that the peptides, which contain proline residue in the binding motifs, can bind to CypA with both *cis* and *trans* prolines maintaining similar interactions between the peptides and CypA(*34*).

In the study of PRS peptides with GYF domain, a register shift motion of the peptides was found for wild type and mutated complexes(*58*) (see Figure 3): Pro5 and Pro6 of the peptide inserted into the binding pocket instead of Pro6 and Pro7 in the original binding modes. Although all four prolines in the peptide are rotated clockwise when viewed from the C to the N terminus, interestingly, the orientations of the remaining residues were kept and show only a slight translation toward the C

terminus. Therefore, all interactions between the peptides and the domain (e.g. electrostatic attractions, hydrophobic interactions and intermolecular hydrogen bonds) were kept. This observation indicates an additional alternative binding mode of the peptides due to their three-fold symmetry around the helical axis (rather than the two-fold rotational symmetry along the helical axis).

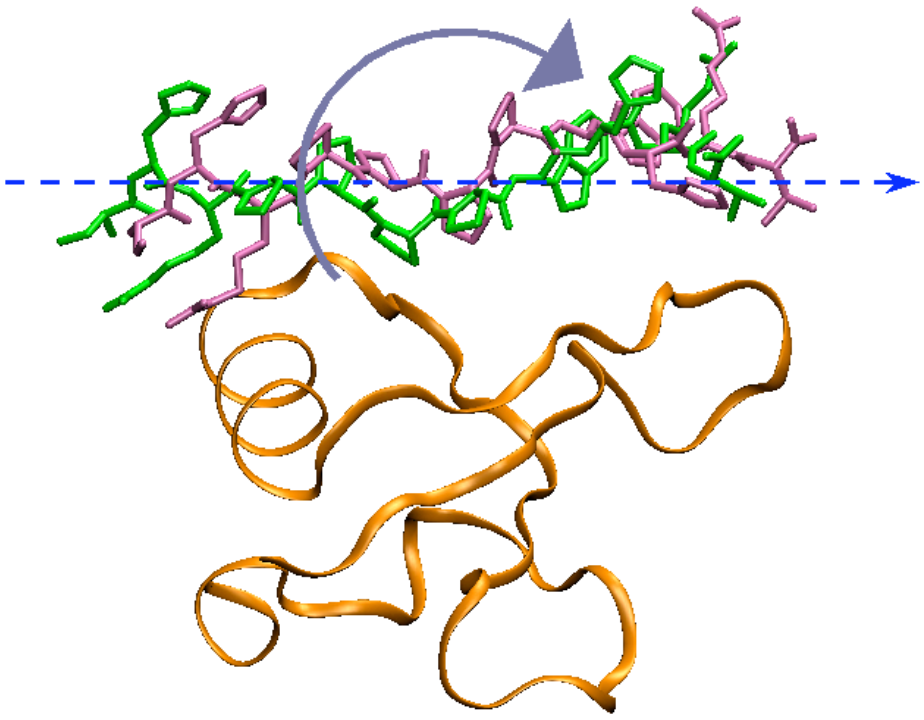


Figure 3: The translation and rotation motions of the peptide between the two binding modes of the PRS ligand (shown as sticks) binding to the GYF domain (shown as ribbons).

What is the function of these different binding modes? Providing two alternative binding modes for a peptide should, theoretically, provide a small additional stability for the bound conformation due to the larger number of states accessible inside the minimum energy well of the bound state. Therefore, this 'screw-like' rotation-translation motion or the transition between different binding modes can decrease the entropic penalty of the binding without affecting the specificity. For the "shift in register" transition between different binding modes of the PRS-GYF system, we suggested an additional function that is related to the binding mechanism: the peptides may bind or leave the binding interface on the recognition domain by this "screw like" motion along the interface. Such screw-like motions may allow for a kinetically favorable binding process by "stripping off water molecules" upon binding and/or unbinding. Furthermore, these sites might act as delocalized anchors within protein associations that rely on fast structural rearrangements within the context of eukaryotic signal transduction(*58*).

4.6 Conclusions and Perspectives in Systems Biology

Proline-rich sequences and their recognition domains are of particular biological importance in signal transduction and complex assembling. The special geometric and chemical properties of proline and PRS make the binding of PRS to the recognition domains rapid and week. In the systems biology point of view, these systems have important roles as mediators in the protein-protein interaction networks of cells (see below). However, recent evidence supports the notion that these interactions are not simple on/off reactions but may be fine-tuned and/or regulated by delicate conformational transitions.

Currently, one of the greatest challenges facing cellular proteomics is to understand the roles of thousands of proteins acting as principal components of a cell, and how they interact to create this complex but organized "machine". This network of interactions, also termed the "interactome", is only one part of the cellular network that also includes the gene regulation network, the metabolic network, the functional network and so on(*60-66*). We argue that for protein-protein interaction networks, three-dimensional structural information is essential for the correct and meaningful establishment of the networks(*67*). The reason is that many proteins interact with each other via extended surfaces composed of 10-30 residues that may be far apart in the sequences. In these cases it is hardly possible to map the interaction just by matching their primary sequences(*60*). However, much work is needed to determine the precise three-dimensional structures of thousands of large protein complexes. Promising steps in this direction are either based on combining results from pair wise docking of rigid proteins(*68*) or on combining experimental information with bioinformatics approaches(*67*). On the other hand, the PRS discussed in this review are mostly short, extended peptides. For this case, both experimental methods (e.g. phage display or yeast two-hybrid) and computational approaches (e.g. pattern matching or flexible docking + refinement) can lead to reasonably accurate interaction data. For example, Cesareni and coworkers studied the binding of PRS with SH3 domains by a combined experimental and theoretical methods(*69*) and established an interaction network between different PRS and SH3 domains.

The weak but rapid binding of PRS to recognition domains makes the system an ideal object of developing protein function networks or protein function predictions(*70, 71*). In this review, we pointed out the importance of accounting for the multiplicity of interactions between PRS and the recognition domains and modifications of protein functions that depend on the isomerization of proline residues. Similar considerations may apply for peptide substrates of protein kinases as well. Here, we mention in particular the inhibitor of cAMP-dependent protein kinase, PKI, which is partly unstructured in the unbound form and folds upon binding to cAPK.

References

1. Kay, B. K., Williamson, M. P., and Sudol, P. (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains, *Faseb J. 14*, 231-241.
2. Sicheri, F., Moarefi, I., and Kuriyan, J. (1997) Crystal structure of the Src family tyrosine kinase Hck, *Nature 385*, 602-609.
3. Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S. B., Gong, F. C., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J. Y., Li, Z. Y., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. Q. H., Zhong, F., Zhong, W. Y., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. (2000) Comparative genomics of the eukaryotes, *Science 287*, 2204-2215.
4. Zarrinpar, A., Bhattacharyya, R. P., and Lim, W. A. (2003) The structure and function of proline recognition domains, *Sci. STKE. RE8*.
5. Carlsson, L., Nystrom, L. E., Sundkvist, I., Markey, F., and Lindberg, U. (1977) Actin polymerizability is influenced by profilin, a low molecular weight protein in non-muscle cells, *J. Mol. Biol. 115*, 465-483.
6. Mayer, B. J., Hamaguchi, M., and Hanafusa, H. (1988) A novel viral oncogene with structural similarity to phospholipase C, *Nature 332*, 272-275.
7. Stahl, M. L., Ferenz, C. R., Kelleher, K. L., Kriz, R. W., and Knopf, J. L. (1988) Sequence similarity of phospholipase C with the non-catalytic region of src, *Nature 332*, 269-272.
8. Bork, P., and Sudol, M. (1994) The Ww Domain - a Signaling Site in Dystrophin, *Trends Biochem. Sci. 19*, 531-533.
9. Niebuhr, K., Ebel, F., Frank, R., Reinhard, M., Domann, E., Carl, U. D., Walter, U., Gertler, F. B., Wehland, J., and Chakraborty, T. (1997) Novel proline-rich motif present in ActA of Listeria monocytogenes and cytoskeletal proteins is the ligand for the EVH1 domain, a protein module present in the Ena/VASP family, *EMBO J. 16*, 5433-5444.
10. Nishizawa, K., Freund, C., Li, J., Wagner, G., and Reinherz, E. L. (1998) Identification of a proline-binding motif regulating CD2- triggered T lymphocyte activation, *Proc. Natl. Acad. Sci. U. S. A. 95*, 14897-14902.
11. Freund, C., Dotsch, V., Nishizawa, K., Reinherz, E. L., and Wagner, G. (1999) The GYF domain is a novel structural fold that is involved in lymphoid signaling through proline-rich sequences, *Nat. Struct. Biol. 6*, 656-660.
12. Pornillos, O., Alam, S. L., Davis, D. R., and Sundquist, W. I. (2002) Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein, *Nat. Struct. Biol. 9*, 812-817.
13. Sancho, E., Vila, M. R., Sanchez-Pulido, L., Lozano, J. J., Paciucci, R., Nadal, M., Fox, M., Harvey, C., Bercovich, B., Loukili, N., Ciechanover, A., Lin, S. L., Sanz, F., Estivill, X., Valencia, A., and Thomson, T. M. (1998) Role of UEV-1, an inactive variant of the E2 ubiquitin- conjugating enzymes, in in vitro differentiation and cell cycle behavior of HT-29-M6 intestinal mucosecretory cells, *Mol. Cell. Biol. 18*, 576-589.
14. Myllyharju, J., and Kivirikko, K. I. (1999) Identification of a novel proline-rich peptide-binding domain in prolyl 4-hydroxylase, *EMBO J. 18*, 306-312.
15. Ball, L. J., Jarchau, T., Oschkinat, H., and Walter, U. (2002) EVH 1 domains: structure, function and interactions, *FEBS Letters 513*, 45-52.
16. Macias, M. J., Wiesner, S., and Sudol, M. (2002) WW and SH3 domains, two different scaffolds to recognize proline-rich ligands, *FEBS Letters 513*, 30-37.
17. Mayer, B. J. (2001) SH3 domains: complexity in moderation, *J. Cell. Sci. 114*, 1253-1263.

18. Eckert, B., Martin, A., Balbach, J., and Schmid, F. X. (2005) Prolyl isomerization as a molecular timer in phage infection, *Nat. Struct. Mol. Biol. 12*, 619-623.
19. Tanaka, S., and Scheraga, H. A. (1975) Calculation of the characteristic ratio of randomly coiled poly(L-proline), *Macromolecules 8*, 623-631.
20. Steinberg, I. Z., Harrington, W. F., Berger, A., Sela, M., and Katchalski, E. (1960) The configurational changes of poly-L-proline in solution, *J. Am. Chem. Soc. 82*, 5263-5279.
21. Deber, C. M., Bovey, F. A., Carver, J. P., and Blout, E. R. (1970) Nuclear magnetic resonance evidence for cis-peptide bonds in proline oligomers, *J. Am. Chem. Soc. 92*, 6191-6198.
22. Mattice, W. L., and Mandelkem, L. (1971) Conformational properties of poly-L-proline form II in dilute solution., *J. Am. Chem. Soc. 93*, 1769-1777.
23. Vila, J. A., Baldoni, H. A., Ripoll, D. R., Ghosh, A., and Scheraga, H. A. (2004) Polyproline II helix conformation in a proline-rich environment: A theoretical study, *Biophys. J. 86*, 731-742.
24. Feng, S. B., Chen, J. K., Yu, H. T., Simon, J. A., and Schreiber, S. L. (1994) 2 Binding Orientations for Peptides to the Src Sh3 Domain - Development of a General-Model for Sh3-Ligand Interactions, *Science 266*, 1241-1247.
25. Schuler, B., Lipman, E. A., Steinbach, P. J., Kumke, M., and Eaton, W. A. (2005) Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence, *Proc. Natl. Acad. Sci. U. S. A. 102*, 2754-2759.
26. Adzhubei, A. A., and Sternberg, M. J. E. (1993) Left-handed Polyproline II Helices Commonly Occur in Globular Proteins, *J. Mol. Biol. 229*, 472-493.
27. Schiene-Fischer, C., and Yu, C. (2001) Receptor accessory folding helper enzymes: the functional role of peptidyl prolyl cis/trans isomerases, *FEBS Letters 495*, 1-6.
28. Fischer, G. (1994) Peptidyl-Prolyl Cis/Trans Isomerases and Their Effectors, *Angew. Chem. Int. Edit. 33*, 1415-1436.
29. Rucker, A. L., and Creamer, T. P. (2002) Polyproline II helical structure in protein unfolded states: Lysine peptides revisited, *Protein Sci. 11*, 980-985.
30. Mezei, M., Fleming, P. J., Srinivasan, R., and Rose, G. D. (2004) Polyproline II helix is the preferred conformation for unfolded polyalanine in water, *Proteins 55*, 502-507.
31. Kentsis, A., Mezei, M., Gindin, T., and Osman, R. (2004) Unfolded state of polyalanine is a segmented polyproline II helix, *Proteins-Structure Function and Bioinformatics 55*, 493-501.
32. Siligardi, G., and Drake, A. F. (1995) The Importance of Extended Conformations and, in Particular, the P-Ii Conformation for the Molecular Recognition of Peptides, *Biopolymers 37*, 281-292.
33. Eisenmesser, E. Z., Bosco, D. A., Akke, M., and Kern, D. (2002) Enzyme dynamics during catalysis, *Science 295*, 1520-1523.
34. Piotukh, K., Gu, W., Kofler, M., Labudde, D., Helms, V., and Freund, C. (2005) Cyclophilin a binds to linear peptide motifs containing a consensus that is present in many human proteins, *J. Biol. Chem. 280*, 23668-23674.
35. Mallis, R. J., Brazin, K. N., Fulton, D. B., and Andreotti, A. H. (2002) Structural characterization of a proline-driven conformational switch within the Itk SH2 domain, *Nat. Struct. Biol. 9*, 900-905.
36. Brazin, K. N., Mallis, R. J., Fulton, D. B., and Andreotti, A. H. (2002) Regulation of the tyrosine kinase Itk by the peptidyl-prolyl isomerase cyclophilin A, *Proc. Natl. Acad. Sci. U. S. A. 99*, 1899-1904.
37. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, *J. Mol. Biol. 293*, 321-331.
38. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. R., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C. H., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, M., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J. Mol. Graph. 19*, 26-59.
39. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity linking function and disorder, *Nat. Biotechnol. 19*, 805-806.
40. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54-60.
41. Verkhivker, G. M., Bouzida, D., Gehlhaar, D. K., Rejto, P. A., Freer, S. T., and Rose, P. W. (2003) Simulating disorder-order transitions in molecular recognition of unstructured proteins: Where folding meets binding, *Proc. Natl. Acad. Sci. U. S. A. 100*, 5148-5153.

42. Parker, D., Rivera, M., Zor, T., Henrion-Caude, A., Radhakrishnan, I., Kumar, A., Shapiro, L. H., Wright, P. E., Montminy, M., and Brindle, P. K. (1999) Role of secondary structure in discrimination between constitutive and inducible activators, *Mol. Cell. Biol. 19*, 5601-5607.
43. Hamburger, J. B., Ferreon, J. C., Whitten, S. T., and Hilser, V. J. (2004) Thermodynamic Mechanism and Consequences of the Polyproline II (P(II)) Structural Bias in the Denatured States of Proteins, *Biochemistry 43*, 9790-9799.
44. Blanco, F. J., Rivas, G., and Serrano, L. (1994) A Short Linear Peptide That Folds into a Native Stable Beta- Hairpin in Aqueous-Solution, *Nat. Struct. Biol. 1*, 584-590.
45. Rucker, A. L., Pager, C. T., Campbell, M. N., Qualls, J. E., and Creamer, T. P. (2003) Host-guest scale of left-handed polyproline II helix formation, *Proteins 53*, 68-75.
46. Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F., and Creamer, T. P. (2001) Host-guest study of left-handed polyproline II helix formation, *Biochemistry 40*, 14376-14383.
47. Williamson, M. P. (1994) The Structure and Function of Proline-Rich Regions in Proteins, *Biochem. J. 297*, 249-260.
48. Shi, Z. S., Olson, C. A., Rose, G. D., Baldwin, R. L., and Kallenbach, N. R. (2002) Polyproline II structure in a sequence of seven alanine residues, *Proc. Natl. Acad. Sci. U. S. A. 99*, 9190-9195.
49. Woody, R. W. (1992) Circular dichroism and conformation of unordered polypeptides, *Adv. Biophys. Chem. 2*, 37-79.
50. Tiffany, M. L., and Krimm, S. (1972) Effect of temperature on the circular dichroism spectra of polypeptides in the extended state, *Biopolymers 11*, 2309-2316.
51. Tiffany, M. L., and Krimm, S. (1968) Circular dichroism of poly-L-proline in an unordered conformation, *Biopolymers 6*, 1767-1770.
52. Asher, S. A., Mikhonin, A. V., and Bykov, S. (2004) UV Raman demonstrates that alpha-helical polyalanine peptides melt to polyproline II conformations, *J. Am. Chem. Soc. 126*, 8433-8440.
53. Chellgren, B. W., and Creamer, T. P. (2004) Short sequences of non-proline residues can adopt the polyproline II helical conformation, *Biochemistry 43*, 5864-5869.
54. Creamer, T. P. (1998) Left-handed polyproline II helix formation is (very) locally driven, *Proteins 33*, 218-226.
55. Sreerama, N., and Woody, R. W. (1999) Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(Pro)II conformations, *Proteins 36*, 400-406.
56. Pappu, R. V., and Rose, G. D. (2002) A simple model for polyproline II structure in unfolded states of alanine-based peptides, *Protein Sci. 11*, 2437-2455.
57. Stapley, B. J., and Creamer, T. P. (1999) A survey of left-handed polyproline II helices, *Protein Sci. 8*, 587-595.
58. Gu, W., Kofler, M., Antes, I., Freund, C., and Helms, V. (2005) Alternative binding modes of proline-rich peptides binding to the GYF domain, *Biochemistry 44*, 6404-6415.
59. Mahoney, N. M., Rozwarski, D. A., Fedorov, E., Fedorov, A. A., and Almo, S. C. (1999) Profilin binds proline-rich ligands in two distinct amide backbone orientations, *Nat. Struct. Biol. 6*, 666-671.
60. Castagnoli, L., Costantini, A., Dall'Armi, C., Gonfloni, S., Montecchi-Palazzi, L., Panni, S., Paoluzi, S., Santonico, E., and Cesareni, G. (2004) Selectivity and promiscuity in the interaction network mediated by protein recognition modules, *FEBS Letters 567*, 74-79.
61. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. U. S. A. 98*, 4569-4574.
62. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature 403*, 623-627.
63. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T.,

Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature 415*, 180-183.

64. Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature 415*, 141-147.
65. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003) A protein interaction map of Drosophila melanogaster, *Science 302*, 1727-1736.
66. Li, S. M., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q. R., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H. Y., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004) A map of the interactome network of the metazoan C-elegans, *Science 303*, 540-543.
67. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004) Structure-based assembly of protein complexes in yeast, *Science 303*, 2026-2029.
68. Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H. J. (2005) Prediction of Multimolecular Assemblies by Multiple Docking, *J. Mol. Biol. 349*, 435-447.
69. Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W. V., Fields, S., Boone, C., and Cesareni, G. (2002) A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules, *Science 295*, 321-324.
70. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps, *Bioinformatics 21*, i302-310.
71. Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004) A probabilistic functional network of yeast genes, *Science 306*, 1555-1558.

Chapter 5

Are solvation free energies of homogeneous helical peptides additive?

(published in *J. Phys. Chem. B*, **109**, 19000-19007 (2005))

5.1 Summary

We investigated the additivity of the solvation free energy of amino acids in homogeneous helices of different length in water and in chloroform. Solvation free energies were computed by multi-configuration thermodynamic integration (MCTI) involving extended molecular dynamics simulations and by applying the Generalized-Born Surface Area (GBSA) solvation model to static helix geometries. The investigation focused on homogenous peptides composed of uncharged amino acids, where the backbone atoms are kept fixed in ideal helical conformation. We found non-linearity especially for short peptides, which is opposing a simple treatment of the interaction of amino acids with their surroundings. For homogeneous peptides longer than 5 residues, the results from both methods are in quite good agreement and solvation energies are to a good extent additive.

5.2 Introduction

It has been well recognized that solvation effects play a crucial role in almost every process in molecular biology, for example in protein folding and the molecular recognition among proteins or for the aggregation of transmembrane helices (*1–8*). All these processes go along with the transfer of a solute, mostly amino acids or proteins, between a polar solvent with a high dielectric constant and a non-polar medium. During transfer, a set of non-covalent contacts is formed or broken within the solute molecules and between solute and solvent molecules. The accurate description of solvation effects is therefore an essential part of any systematic approach aiming at contributing to the understanding of such processes.

Over the past decades many experimental studies have addressed the solvation properties of amino acids as well as of peptides (*9–14*). However, due to the very different physico-chemical properties among the 20 naturally occurring amino acids the respective experimental techniques are facing significant challenges. Theoretical modelling of biological systems is thereby highly desirable to complement experimental studies (*3, 15–18*).

When applying computational methods for deriving solvation free energies of peptides or proteins, one constantly faces the dilemma of achieving both physical accuracy and computational efficiency. The most reliable theoretical method available today are free energy calculations that have been thoroughly refined during the 1990'ies allowing for systematic studies of solvation properties. Two variants of these are free energy pertubation (*19*) and thermodynamic integration (*20*). Recent studies on the hydration free energies of amino acid side-chain analogs (*15–18*) using multi-configuration thermodynamic integration (MCTI) (*20*) with separation-shifted potential scaling (*21, 22*) achieved very satisfactory agreement with experimental studies. However, this method requires an explicit representation of solvent molecules and the results critically depend on how complete the relevant parts of the conformational space were sampled. These requirements make the method computationally very expensive or even prohibitive when being applied to large systems like proteins.

Implicit solvent models reduce the explicit interactions between solute and solvent to a mean field property that only relies on the solute conformation (*3, 23, 24*). Therefore they are currently heavily used in areas ranging from protein structure prediction (*25–30*), protein folding (*4–8, 31–35*), and modelling protein-protein/ligand interactions (*4, 36–39*). All these implicit solvent models assume, either in part (*23, 24, 40–43*) or completely (*44–52*), that solvation free energy contributions due to neighboring segments are additive. Whereas additivity is certainly not fulfilled for charged amino acids, this assumption is based on the idea that the interactions of polar and non-polar side chains affecting the solvent structure are of short-range nature. Supporting evidence comes from an experimental study of solubilities of the peptide backbone unit in various solvents (*14*). There, an additivity of backbone transfer free energies was found. On the other hand, a theoretical study of the formation of secondary structure observed non-additivity for the free energies of the formation of short α-helices using the Finite Difference Poisson-Boltzmann method (*53*). It appears that this implicit assumption of solvation free energies being additive has not comprehensively been tested so far.

Nowadays, parametric studies of implicit solvent models are focusing on closely matching the data from experimental and explicit solvent simulations of small molecules (*43, 54, 55*). However, if solvation free energies of neighboring segments are not additive, would it still be suitable to extend the parameters derived from data for small molecules to large systems, or the other way round, to apply models parametrized for large systems to small molecules?

What is an appropriate method to answer these questions? Experimental results have problems when it comes to decomposing results into sequence dependent and conformation dependent contributions. Another problem concerns the solubility of peptides that often requires addition of blocking groups. Fortunately, such issues are less of a problem in theoretical studies. As mentioned before, the most reliable theoretical method available is multi-configuration thermodynamic integration (MCTI). However, due to the large computational efforts involved, applications of MCTI to the computation of solvation free energies were so far restricted to single amino acids. This study is the first attempt to tackle poly-peptide systems up to 9-residues in length. Therefore, an important test was to compare the results from MCTI

calculations with GBSA, one of the most popular and efficient implicit solvent models.

As an extension to our previous work (*15*) and in order to combine this work with an ongoing pro ject in our group designing a residue scale force field for the structure prediction of transmembrane proteins (*56*) we chose homogenous α-helical peptides of different length as model systems for this study. This choice was motivated by the following considerations: (i) restricting the peptide backbone to a given conformation facilitates the sampling during the simulations, (ii) focussing on homogeneous α-helical peptides keeps the sequence depen- dent contributions obvious and understandable and (iii) by comparing the results for different types of amino acid residues, one may attempt to dissect the backbone contributions from the side-chain contributions. We note, though, that backbone and side-chain contributions are commonly interdependent and a true separation is not possible in a strict sense.

We investigated the additivity of the solvation free energy of amino acids in homogeneous helices of different length in water and in chloroform. Solvation free energies were computed by multi-configuration thermodynamic integration (MCTI) involving extended molecular dynamics simulations and by applying the Generalized-Born Surface Area (GBSA) solvation model to static helix geometries. The investigation focused on homogenous peptides composed of uncharged amino acids, where the backbone atoms are kept fixed in ideal helical conformation. We found non-linearity especially for short peptides, which is opposing a simple treatment of the interaction of amino acids with their surroundings. For homogeneous peptides longer than 5 residues, the results from both methods are in quite good agreement and solvation energies are to a good extent additive

5.3 Materials and Methods

5.3.1 Molecular dynamics simulations

This study addresses the solvation properties of homogenous α-helices composed of un- charged amino acids. The coordinates of such α-helices of length 5 were modelled using the TINKER (*57*) package. Each peptide $(X)_n$ is flanked by two glycine residues of the form Gly-$(X)_n$-Gly (see Figure 1). The atoms of the two flanking glycine residues were treated as ”dummy” atoms (see below under ”free energy calculations”). The systems are named GX5G (X refers to the one letter code of amino acids). For the cases of alanine and asparagine we also investigated their homogenous *α*-helices of length 2, 3, 4, 5, 6, and 9 (named GXnG, where *n* refers to the number of residues). The dihedral angles of the peptide backbone were set to the values of an ideal α-helix ($\varphi = -58^\circ$, $\psi = -47^\circ$). For comparison, two five-residue-long homogenous peptides were also modelled with extended conformation ($\varphi = -135^\circ$, $\psi = 135^\circ$) (named GA5GST and GN5GST). During the simulations, all backbone atoms were kept fixed in their starting geometry because we want to investigate the effect of the helical geometry. We note, of course, that an α-helix may not be the preferred conformation in solution for some of the investigated sequences. Molecular dynamics simulations were performed both in chloroform and water as solvent.

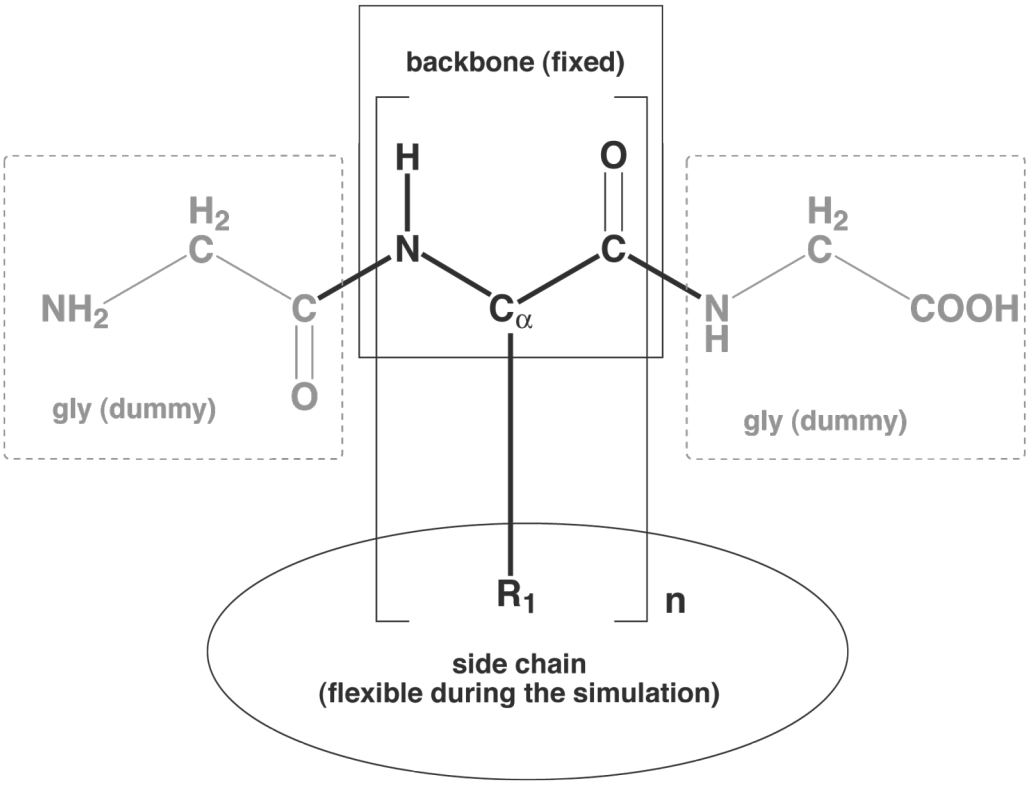


Figure 1: The structure of the system used for the MD simulations and the MCTI calculations. The atoms of the central residues are shown in bold and dummy atoms are colored in grey.

All simulations were carried out employing the NWChem 4.5 package (*58*) with the AMBER99 force field (*59*). The atomic charges of the chloroform model were -0.3847 e for the carbon atom, 0.2659 e for the hydrogen atom and 0.0396 e for the chlorine atoms, respectively. The molecules were solvated in cubic boxes of 4.0 nm side length, using chloroform or TIP3P Water molecules (*60*), respectively, with an initial minimum distance of at least 1.3nm between the boundaries of the box and the nearest solute atom (excluding dummy atoms). All coordinate sets were first optimized by 500 steps of steepest-descent energy minimization. The solvent and modeled residues were then relaxed during a 1ns molecular dynamics (MD) simulation at 300K prior to the free energy calculation. The SHAKE procedure (*61*) was applied to constrain all bonds that contain hydrogen atoms. The time step of the simulations was 2 fs throughout. Non-bonded interactions were treated using a cutoff of 1.2 nm. The temperature and pressure were maintained by weak coupling to an external bath in all simulations (*62*). For the simulations in chloroform the pressure coupling time was set to 5.0 ps and the isothermal compressibility was set to $9.98*10^{-10}m^2N^{-1}$. For the simulations in water the coupling time and compressibility were 0.5 ps and $4.53*10^{-10}m^2N^{-1}$.

5.3.2 Free energy calculations

The solvation free energies of all peptides were calculated according to the following thermodynamic cycles:

$$\Delta G_{solv,peptide} = \Delta G_{peptide \longrightarrow dummy,vacuum} - \Delta G_{peptide \longrightarrow dummy,solvent} + \Delta G_{solv,dummy} \quad (1)$$

$\Delta G_{peptide \longrightarrow dummy,vacuum}$ and $\Delta G_{peptide \longrightarrow dummy,solvent}$ are the free energy differences for switching off the solute-solvent non-bonded interactions (van der Waals interactions and electrostatic interactions) while keeping their bonded interactions and atomic masses unchanged. $\Delta G_{solv,dummy}$, the free energy change for transferring dummy atoms from vacuum into solvent, is zero by definition. Solvation free energies calculated by such thermodynamic cycles are concentration-independent (*63*). The detailed description of free energy calculation (MCTI) can be found in Chapter 1 of this thesis.

The non-bonded interactions between the initial state and the final state are interpolated by a separation-shifted potential scaling (21) using $\delta = 0.075$ nm to avoid the well-known origin singularities. In our study, separate simulations were performed at 21 equally spaced points of λ from $\lambda = 0$ to $\lambda = 1$. At each point, the system was first equilibrated for 200 ps and data were collected during further 200 ps of simulation. The van der Waals and columbic terms were turned off simultaneously. A similar protocol was previously used to compute solvation free energies of small model substances.[66] There, ΔG_{hydr} could be reliably computed with statistical errors of ≤ 1.5 kJ mol^{-1} . The protocol is also similar to the recent studies of Gu et al. (*15*) Villa and Mark (*16*) Shirts et al. (*17*) and of Deng and Roux (*18*) to compute ΔG_{hydr} for amino acid side chain analogs. The convergence of the derivatives of the Hamiltonian with respect to λ was monitored for all individual windows and showed smooth behavior for all computed values (data not shown).

The MCTI calculations were performed under constant pressure conditions. Consequently, upon mutating the peptide into a dummy molecule, the volume of the simulation box shrank. When computing free energies of solvation, the computed values are concentration-independent as noted previously (*63*). Because this study reports for the first time the application of MCTI to remove an entire peptide, the volume of the simulation box was checked during the MCTI calculation for GN9G, since GN9G has the largest solute volume in this study. In the first window, the volume of the simulation box is 64.0 ± 0.2 nm^3. In the last window of the MCTI calculation, in which the solute becomes invisible to the solvent molecules, the volume of the simulation box is 62.9 ± 0.2 nm^3. The observed volume difference of 1.1 ± 0.3 nm^3 is in reasonable agreement with the volume of the simulated Asparagine 9-mer of 0.8 nm^3 (contact/reentrant volume, calculated by TINKER with a probe radius of 0.14 nm). The entropy changes due to modifying the water-peptide interactions are all taken into account by the MCTI method. As long as the peptide atoms are interacting with the solvent molecules, the volume occupied by the peptide is inaccessible to the solvent. Upon tuning the peptide interactions off, the volume of the simulation box shrinks by an amount comparable to the volume of the peptide as required by the condition of constant density. Therefore, the translational entropy of the bulk waters in the box remains unchanged.

For comparison, one of the most popular implicit solvent models, the Generalized Born Surface Area model (GBSA) (*23*) implemented in the TINKER package (*57*), was used to calculate the solvation free energies of homogenous *α*-helices from a length of one up to twenty residues. All peptides are capped with ACE-(CH_3C=O) at

the N-terminus and -NH_2 at the C-terminus. The contributions of the capping groups are subtracted from the total solvation free energies.

5.4 Results

Residue	$\Delta G^{(5)}_{H_2O}$ (kJ/mol)	Ratio *c*	$\Delta G^{(5)}_{H_2O}$ (kJ/mol)	Ratio *c*
GY5G	200.5±4.5	0.91	189.7±3.5	0.77
GW5G	180.7±4.8	0.85	194.6±3.7	0.71
GV5G	100.3±3.8	1.06	117.9±2.6	0.75
GC5G	139.8±3.4	0.93	142.0±2.5	1.04
GF5G	115.7±4.4	1.01	161.3±3.1	0.78
GG5G	156.1±2.8	1.27	102.0±1.9	1.05
GI5G	92.7±4.0	1.01	125.8±2.8	0.74
GL5G	88.2±4.0	1.08	125.1±2.9	0.78
GM5G	111.1±4.1	1.06	147.0±2.8	0.91
GN5G	295.2±3.6	0.89	156.7±2.8	0.61
GT5G	140.4±3.9	0.86	116.9±2.7	0.83
GS5G	183.8±3.6	0.95	115.3±2.4	0.80
GA5G	128.4±3.0	1.28	106.3±2.2	0.97
GA5GST	113.1±1.1	1.13	110.2±2.1	1.01
GN5GST	262.2±1.6	0.79	152.4±2.7	0.59

Table 1 Solvation free energies of five-residue-long homogenous peptides calculated with MCTI.

Solvation free energies of homogeneous helical peptides were computed from molecular dynamics simulations where during the simulation the interactions between solute and solvent are progressively switched off (see the methods section). To derive the solvation free energies of the peptides in water or chloroform, respectively, thermodynamic cycles are constructed where the vacuum-values are subtracted from those in water or in chloroform. Table 1 lists the values of all peptides of length 5 (both helical and extended conformations) in water and in chloroform as well as the ratio

$$c = \frac{\Delta G^{(n)}}{n \cdot \Delta G^{(1)}} \tag{2}$$

This ratio is the solvation free energy of the whole peptide ($\Delta G^{(n)}$) divided by the solvation free energy of the single amino acids ($\Delta G^{(1)}$ taken from (15)) multiplied by the number of residues (n). In this case $n = 5$.

In a couple of cases a rather surprising result is obtained: the ratio c is larger than one. This means that the solvation free energy of the entire peptide is larger than the sum of the single values. Subsequently, we refer to this effect as ”super-unity”. If one considers applying simple residue scaled models to the solvation free energy of peptides, one may formulate

$$\Delta G^{(n)} = \sum_i S_i \cdot \Delta G^{(1)} \tag{3}$$

Here, $\Delta G_i^{(1)}$ should be the solvation free energy of residue i as a single residue and S_i is the ratio of the solvent accessible surface area (SASA) of that residue in the peptide context in a particular conformation (here: helical) relative to the SASA value of the isolated residue ($S_i \leq 1$). When combining equations (2) and (3) it follows that the ratio c must be smaller than one and more or less constant, which implies a linear behaviour of the solvation free energy. We conclude that models that are purely based on SASA terms, fully depend on the linearity of solvation free energies.

To investigate this effect more detailed for two selected systems, we performed MCTI calculations for homogeneous α-helical peptides of 2 to 6 and 9 residues length. Alanine and asparagine were selected for the calculations as examples of non-polar and polar residues. Due to their different size we also expected different contributions from the residue backbone and the side-chain. Water and chloroform were selected as solvent environment to study the effects of different dielectric constants and different sizes of the solvent molecules. In our previous study on individual amino acid ΔG_{solv}, the two solvents yielded results in very good agreement with experimental data. In order to compare with available implicit solvent models, solvation free energies of peptides with length of 1 to 20 residues were computed by the GBSA model as well.

5.4.1 MCTI in water

Figure 2 shows the results for poly-Ala (GAnG, n refers to the number of residues) and poly-Asn (GNnG) of different lengths in aqueous solution using MCTI. For comparison, corresponding results calculated by GBSA are shown as well.

Figure 2 reveals three noteworthy features: First, the curve shapes for GAnG and GNnG are different: a non-linear behavior was found in the GAnG calculation, while the plot for GNnG shows a nearly linear behavior. Second, super-unity ($c > 1$) is observed for both systems but is stronger in GAnG. Third, the results from the two different approaches (MCTI and GBSA) show surprisingly good agreement, especially for GNnG. Sizable differences still exist for GAnG, where the ratio c remains larger than one up to the maximum length of 9 residues in the MCTI calculations, while in the GBSA calculations it reaches a value smaller than one for $n > 3$. For GNnG, the ratio shows a very similar trend in both calculations: it is below one for $n > 3$ and reaches a relatively constant value of 0.8 to 0.9.

In the two cases investigated in extended conformations (GA5GST and GN5GST, see Table 1), the super-unity of the solvation free energies is weaker than in the cases of helical conformation. This indicates that the super-unity is conformationally dependent in water solution.

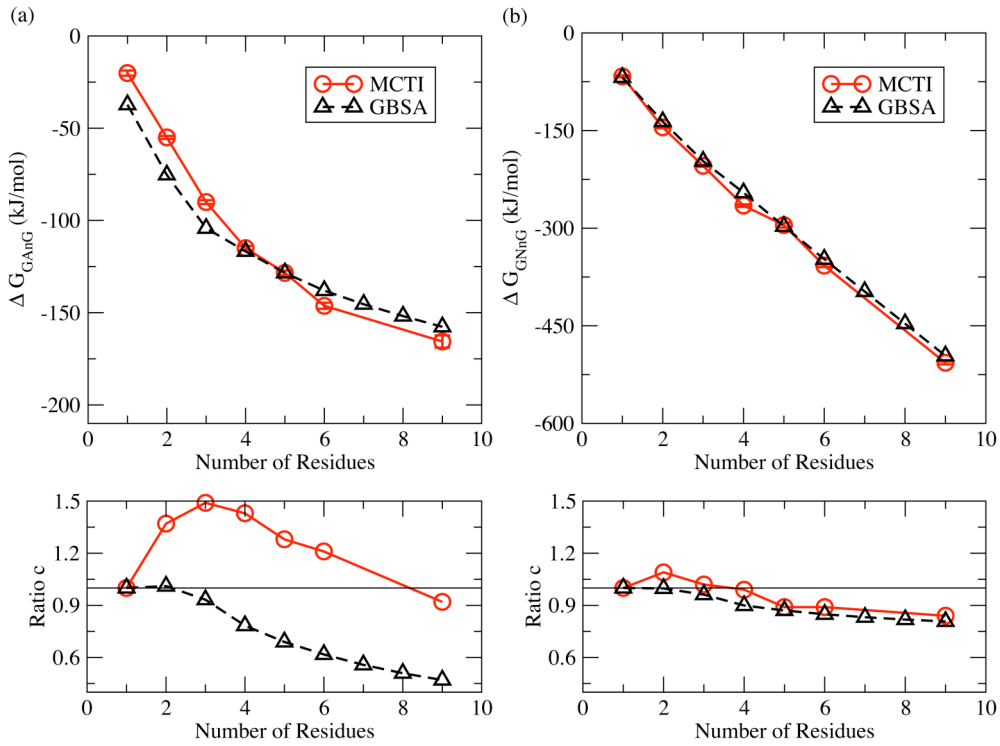


Figure 2: Solvation free energies for poly-alanine-peptides (left) and poly-asparagine-peptides (right) of different length in water from the MCTI and the GBSA calculations.

5.4.2 MCTI in chloroform

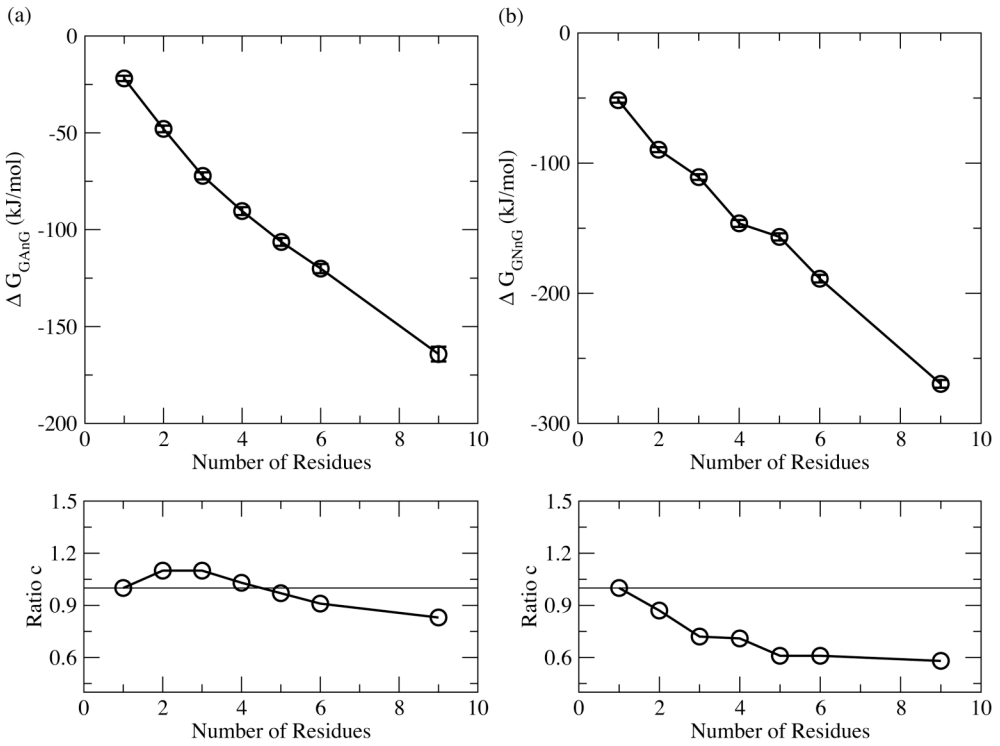


Figure 3: Solvation free energies for poly-alanine-peptides (left) and poly-asparagine-peptides (right) of different length in water from the MCTI calculations.

Figure 3 shows the results for GAnG and GNnG in chloroform solution using MCTI. The results are quite different from those in water as a nearly linear behavior was observed in both systems. The ratio c reaches relatively constant values in both systems. For GAnG it reaches ca. 0.8 for $n > 6$ for the GAnG calculations and the for GNnG it approaches 0.6 for $n \geq 5$. Super-unity was only found in the GAnG calculations for $n < 5$. In the GNnG calculation, no super-unity was found. In this solvent, the results of GA5GST and GN5GST are quite close to those of GA5G and GN5G (see Table 1). The influences due to different backbone conformations are not as large as those in the water solution.

5.4.3 GBSA

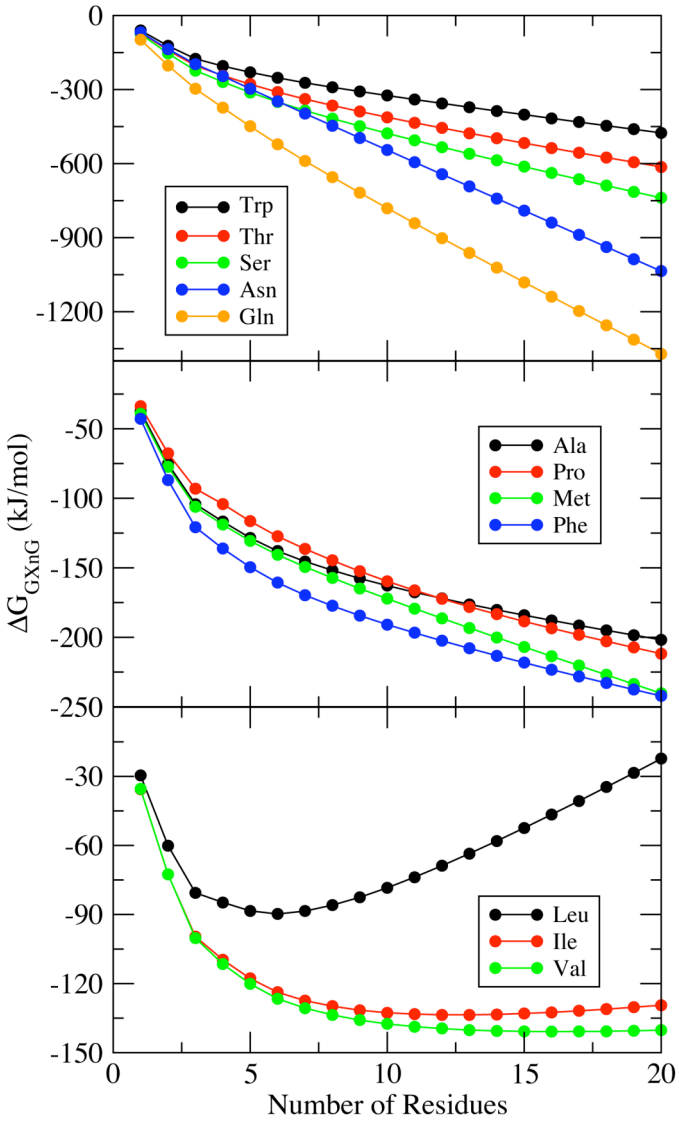


Figure 4: Solvation free energies for homogeneous α-helical peptides of different lengths calculated by the GBSA implicit solvent model.

The solvation free energies of homogenous helical peptides from length of 1 to 20 residues are shown in Figure 4. The peptides can be grouped into three classes according to the properties of the amino acids. The first class comprises the polar amino acids Asn, Ser, Gln, Hid (Histidine with a proton at $N_{\delta 1/\pi}$), Hie (Histidine with a proton at $N_{\varepsilon 2/\tau}$), Thr, Thr, Trp, and Tyr that show a very steep descent and reach values between -400 and -1400 kJ/mol^{-1} for a 20-residue-long peptide. Linear behavior is clearly observed for all members of this class for $n > 5$.

The second class is formed by the non-polar amino acids Ala, Met, Phe, and Pro. They show a clearly non-linear behavior for small peptides until n >10 where they converge into a linear regime. Even though the amino acids in this class are non-polar residues, the solvation free energies still decrease when the number of residues increases.

Ile, Val and Leu, which contain aliphatic side-chains, constitute the third class. The mini- mum solvation free energies reach -140 kJ mol^{-1} to -90 kJ mol^{-1} before reversing their slope. Furthermore, they converge to a linear behavior only for long peptides ($n \geq 10$ for Leu and $n \geq 15$ for Val and Ile). Leu clearly shows strongly opposing contributions. The linear part for long peptides shows that Leu is unfavorable in an aqueous environment. The negative slope for smaller peptides reflects the effect of unsaturated hydrogen bonds as well as other effects (see discussion).

5.5 Discussion

5.5.1 Non-additivity and super-unity

Both methods, the MCTI calculations and GBSA calculations, show non-linearity for short peptides in most cases investigated. Basic considerations show that modeling helical peptides by adding one residue after the other will lead to some discontinuities in the solvation free energies. Figure 5 shows the backbones of peptides of different length (n = 1 to 5). Up to a length of four residues there exist only next neighbors in the same turn and their backbone peptide bonds do not form direct interactions (hydrogen bonds). Therefore, the contributions to the solvation free energy of each residue may be almost independent. From five residues on, there are additionally next-turn-neighbors, which will form inter-molecular hydrogen bonds between backbone atoms. The interactions between the newly added NH group at position n and the surrounding solvent molecules are shielded by the C=O group of residue $n - 4$ because of this inter-molecular hydrogen bond. As a consequence, the interactions between the C=O group at position $n - 4$ and solvent are shielded by the same hydrogen bond as well. This means that from four residues on, the number of ”unsaturated backbone groups”(backbone N-H or C=O that are not involved in intra-peptide hydrogen bonds) does not increase when the helix is extended. Table 2 lists the number of unsaturated backbone groups as a function of n. This number remains

constant when $n \geq 4$.

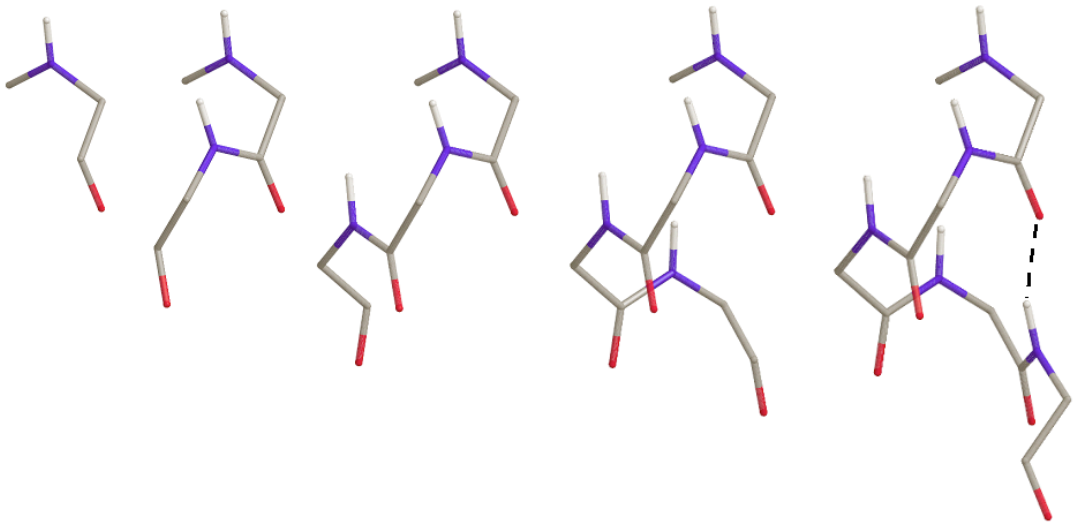


Figure 5: Backbone structures of α−helical peptides of length 1 to 5. Inter-molecular hydrogen bonds are shown in dashed lines.

n	$n_{backbone}$
1	2
2	4
3	6
≥ 4	8

Table 2 Number of unsaturated backbone groups ($n_{backbone}$) as a function of the number of residues (n).

Solvation free energies of organic molecules are commonly decomposed into a non-polar term and a polar term. The non-polar term, which includes the energy cost to form a cavity for the solute in the solvent and to establish van der Waals interactions between the solute and the solvent molecules, is in principle additive with respect to the number of peptide residues n. The polar term includes electrostatic interactions (monopole, dipole and higher multi-poles). This term is most likely not additive per se. In a helical peptide, all dipoles of the backbone point in the same direction and therefore form an overall dipole along the helix axis (*67*). This dipole will align water molecules in the solvation shell around the peptide. Concerning the scaling of this contribution with the peptide length n: the first residue induces orientational polarization of all solvent molecules inside a shell around the backbone. When the length of the helical peptide is increased, eventually all solvent will be orientationally polarized within a cylinder around the helical peptide. The volume of this cylinder grows proportionally to n. Therefore, the electrostatic contribution of the solvation free energy should approach a linear dependence with peptide length for $n \geq$

5 − 10 while it may display non-linearity for shorter peptides. For amino acids with non-polar side-chains, e.g. class 2 and class 3 in the GBSA calculation, the contributions of their backbone groups are the dominant terms in water solution. The discontinuity of the solvation free energies (see Figure 2 and Figure 4) can thus be traced back to the discontinuity of the electrostatic contribution of free backbone groups (see Table 2). Clearly, this behavior is more noticeable in water than in chloroform. Upon increasing the number of residues n the contribution of the dipole term increases linearly whereas the overall contribution of the electrostatic term is limited by the number of unsaturated groups. Therefore, the solvation free energy as a function of n becomes linear and additive for longer peptides (e.g. $n \geq 10$). For amino acids with polar side-chains or very large side-chains, the electrostatic contribution of the side-chains is comparable to that from the backbone groups. Because the number of side-chains and the solvent surface area grows approximately linear with the number of residues, this term is nearly linear. In addition, polar or very large side-chains shield the backbone from the solvent molecules, which somehow weakens the contribution of the backbone. Consequently, the non-additive effect is less pronounced than in non-polar amino acids and linearity or additivity may be observed in shorter peptides as well. Generally speaking, however, the shielding of the backbone by large side-chains should be more important for longer peptides than for shorter peptides. In other cases, e.g. for Leu and Ile, where the solvation free energies increase after a certain inflexion, the geometry of their aliphatic side- chains leads to a stronger shielding of the backbone and the side-chain contribution becomes dominant. Because real systems are composed of a mixture of different amino acid types such deviations from linearity may partially compensate each other.

The same reasoning can also be applied to the effect of different solvent environments. Figure 3 shows that in chloroform nearly linear behavior of peptide solvation free energies was observed for both non-polar and polar amino acids. This is well understandable because in a less polar or non-polar solution, the importance of electrostatic interactions decreases. As a result, the non-polar term, which is in principle additive, becomes dominant in this case.

The ”super-unity” effect observed for short peptides with small and non-polar side-chains (e.g. Ala and Gly) can be explained by this reasoning as well. In such peptides, the side-chain contribution is less important in magnitude. When a residue is added to the peptide, a part of the surrounding water molecules is already adapted to the overall dipole of the helix. This reduces the cost of aligning the nearby solvent molecules compared to solvating an individual residue. Therefore it is more favorable to add an amino acid at the terminus of a short peptide than solvating the first amino acid of this peptide. When the peptide length increases beyond $n = 4$, the involvement of the backbone group in intra-helical hydrogen bonds reduces the contribution of the newly added residues. Thus super-unity does not exist anymore. We note that a 1.2 nm cutoff was applied to all non-bonded interactions in the MCTI simulations for technical reasons. Because the central peptide units will thus feel their electrostatic environment only within a limited range, the use of a cutoff may enhance additivity for long peptides. On the other hands, the MCTI calculations were only performed for systems up to 9 residues long. The calculations of this study about additivity for long peptides are mainly based on the GBSA results, where no cutoff was applied.

5.5.2 Implications from non-additivity

Hydrophobic free energies are commonly derived from experimentally determined solubility of small organic compounds like hydrocarbons ("microscopic") (*68–71*). In prior parameterizations for the hydrophobic effect, a correlation was proposed between the hydrophobic free energy changes and the solvent accessible surface area (*72, 73*). The values derived for transfer from vacuum to water are typically in the order of 5-7 cal mol^{-1} $Å^{-2}$ (*74*). Can the values derived from small molecules easily be transferred to other scales? Our calculations identified different slopes for short peptides and long peptides that are due to mixed electrostatic and hydrophobic contributions. Deriving a parameterization of the hydrophobic effect would require a careful decomposition of these. However, the good agreement between MCTI and GBSA results for longer peptides indicates that the SASA-term in GBSA must work quite well already, yielding a useful parameterization of the hydrophobic effect. Nonetheless, these results indicate that one should apply caution when transferring results that were derived for molecules of different sizes. The respective parameterization should be chosen based on experimental data for the same scale of the problem.

After accepting the consequences of non-additivity of solvation free energies, it may seem that implicit solvent models based on additivity are problematic *perse*. Nevertheless, since in many atomic scaled implicit solvent models, e.g. GBSA, the solvation term is decomposed into polar and non-polar terms, which are individually treated at atomic scale, these models are likely not affected by the phenomenon of non-additivity. Only such models that fully rely on the SASA term may need some improvements. Here, a newly developed model of calculating non-local electrostatics interactions may be helpful (*75*). Because the implicit solvent models at atomic scale are still quite expensive for large-scale problems such as flexible protein-protein docking, assembly of transmembrane helices or protein complexes, we believe that implicit solvent models at residue scale should have very promising applications in those areas.

5.6 Conclusion

The conclusions of the present study are restricted to (i) fully homogenous peptides composed of uncharged amino acids, (ii) that are kept in a frozen backbone helical conformation, (iii) and are fully solvated. Based on the investigated systems, we find:

1. Solvation free energies of peptides of various length were computed by the MCTI and GBSA methodologies. For 5 or more residues the results are in quite good agreement. This observation gives strong support for our strategy of computing ΔG_{hydr} for peptides up to 9 residues from MCTI calculations. However, MCTI and GBSA still show sizable differences for short helices where MCTI should be quite accurate. Thus, it is important to consider molecular details of backbone hydration.

2. Non-additivity is found by both methodologies for peptides shorter than 5 residues. On the other hand, according to the GBSA calculations, additivity appears fulfilled for helices longer than 10 residues. This points towards using caution when transferring SASA parameters that are extracted on the basis of solubility or partition coefficients of small molecules to large systems. The other way round, it may be also problematic to use values that are derived from large systems to small molecules.

3. The design of simplified models, where helices are composed of residue-beads and interactions are modeled additively, appears challenging.

Future work is needed that extends investigations of this type to heterogeneous sequences to see if additivity of solvation energies holds in a general sense.

References

1. Honig, B.; Yang, A. S. (1995), *Adv. Protein Chem.* 46, 27–58.
2. Dill, K. A. (1990), *Biochemistry* 29, 7133–7155.
3. Feig, M.; Brooks, C. L. (2004), *Curr. Opin. Struct. Biol.* 14, 217–224.
4. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. (2004), *J. Med. Chem.* 47, 3032–3047.
5. Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. (2003), *Curr. Opin. Struct. Biol.* 13, 168–174.
6. Pitera, J. W.; Swope, W. (2003), *Proc. Natl. Acad. Sci. USA* 100, 7587–7592.
7. Ohkubo, Y. Z.; Brooks, C. L. (2003), *Proc. Natl. Acad. Sci. USA* 100, 13916–13921.
8. Nymeyer, H.; Garcia, A. E. (2003), *Proc. Natl. Acad. Sci. USA* 100, 13934–13939.
9. Fauchere, J. L.; Pliska, V. Eur. (1983), *J. Med. Chem.* 18, 369–375.
10. Kim, A.; Szoka, F. C. (1992), *Pharma. Res.* 9, 504–514.
11. Radzicka, A.; Wolfenden, R. (1988), *Biochemistry* 27, 1664–1670.
12. Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. (1981), *Biochemistry* 20, 849–855.
13. Wimley, W. C.; Creamer, T. P.; White, S. H. (1996), *Biochemistry* 35, 5109–5124.
14. Auton, M.; Bolen, D. W. (2004), *Biochemistry* 43, 1329–1342.
15. Gu, W.; Rahi, S. J.; Helms, V. (2004), *J. Phys. Chem. B* 108, 5806–5814.
16. Villa, A.; Mark, A. E. (2002), *J. Comput. Chem.* 23, 548–553.
17. Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. (2003), *J. Chem. Phys.* 119, 5740–5761.
18. Deng, Y. Q.; Roux, B. (2004), *J. Phys. Chem. B* 108, 16567–16576.
19. Kollman, P. (1993), *Chem. Rev.* 93, 2395–2417.
20. Straatsma, T. P.; McCammon, J. A. (1991), *J. Chem. Phys.* 95, 1175–1188.
21. Zacharias, M.; Straatsma, T. P.; McCammon, J. A. (1994), *J. Chem. Phys.* 100, 9025–9031.
22. Beutler, T. C.; van Gunsteren, W. F. (1994), *J. Chem. Phys.* 101, 1417–1422.
23. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. (1990), *J. Am. Chem. Soc.* 112, 6127–6129.
24. Constanciel, R.; Contreras, R. (1984), *Theo. Chim. Acta.* 65, 1–11.
25. Dominy, B. N.; Brooks, C. L. (2002), *J. Comput. Chem.* 23, 147–160.
26. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. (2002), *Proteins: Struct. Funct. Genet.*48, 404–422.
27. Feig, M.; Brooks, C. L. (2002), *Proteins: Struct. Funct. Genet.*49, 232–245.
28. Zhu, J.; Zhu, Q. Q.; Shi, Y. Y.; Liu, H. Y. (2003), *Proteins: Struct. Funct. Genet.* 52, 598–608.
29. Forrest, L. R.; Woolf, T. B. (2003), *Proteins: Struct. Funct. Genet.* 52, 492–509.
30. Fiser, A.; Feig, M.; Brooks, C. L.; Sali, A. (2002), *Acc. Chem. Res.* 35, 413–421.
31. He, J. B.; Zhang, Z. Y.; Shi, Y. Y.; Liu, H. Y. (2003), *J. Chem. Phys.* 119, 4005–4017.
32. Zagrovic, B.; Sorin, E. J.; Pande, V. (2001), *J. Mol. Biol.* 313, 151–169.
33. Zhou, R. H. (2003), *Proteins: Struct. Funct. Genet.* 53, 148–161.
34. Suenaga, A. (2003), *J. Mol. Struct-Theo.* 634, 235–241.
35. Liu, Y. X.; Beveridge, D. L. (2002), *Proteins: Struct. Funct. Genet.* 46, 128–146.
36. Donnini, S.; Juffer, A. H. (2004), *J. Comput. Chem.* 25, 393–411.
37. Lazaridis, T. (2003), *Proteins: Struct. Funct. Genet.* 52, 176–192.
38. Mardis, K. L.; Luo, R.; Gilson, M. K. (2001), *J. Mol. Biol.* 309, 507–517.
39. Gohlke, H.; Case, D. A. (2003), *J. Comput. Chem.* 25, 238–250.
40. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. (1997), *J. Phys. Chem. A* 101, 3005–3014.
41. Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. (2003), *J. Comput. Chem.* 24, 1348–1356.
42. Im, W.; Lee, M. S.; Brooks, C. L. (2003), *J. Comput. Chem.* 24, 1691–1702.
43. Zhu, J. A.; Shi, Y. Y.; Liu, H. Y. (2002), *J. Phys. Chem. B* 106, 4844–4853.
44. Gallicchio, E.; Zhang, L. Y.; Levy, R. M. (2002), *J. Comput. Chem.* 23, 517–529.
45. Wesson, L.; Eisenberg, D. (1992), *Protein Sci.* 1, 227–235.
46. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. (1987), *Proc. Natl. Acad. Sci. USA* 84,

3086–3090.
47. Ferrara, P.; Apostolakis, J.; Caflisch, A. (2002), *Proteins: Struct. Funct. Genet.* 46, 24–33.
48. Weiser, J.; Shenkin, P. S.; Still, W. C. (1999), *Biopolymers* 50, 373–380.
49. Hou, T. J.; Qiao, X. B.; Zhang, W.; Xu, X. J. (2002), *J. Phys. Chem. B* 106, 11295–11304.
50. Zhou, H. Y.; Zhou, Y. Q. (2002), *Proteins: Struct. Funct. Genet.* 49, 483–492.
51. Lomize, A. L.; Riebarkh, M. Y.; Pogozheva, I. D. (2002), *Protein Sci.* 11, 1984–2000.
52. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. (2003), *J. Am. Chem. Soc.* 125, 9523–9530.
53. Yang, A. S.; Honig, B. (1995), *J. Mol. Biol.* 252, 351–365.
54. Nina, M.; Beglov, D.; Roux, B. (1997), *J. Phys. Chem. B* 101, 5239–5248.
55. Banavali, N. K.; Roux, B. (2002), *J. Phys. Chem. B* 106, 11026–11035
56. Park, Y.; Elsner, M.; Staritzbichler, R.; Helms, V. (2004), *Proteins: Struct. Funct. Bioinfo.* 57, 577–585.
57. Ren, P. Y.; Ponder, J. W. (2003), *J. Phys. Chem. B* 107, 5933–5947.
58. Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J. L.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. (2001), *Comput. Phys. Commun.* 128, 260–283.
59. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. (1996), *J. Am. Chem. Soc.* 118, 2309–2309.
60. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. (1983), *J. Chem. Phys.* 79, 926–935.
61. Ryckaert, J.; Ciccotti, G.; Berendsen, H. (1977), *J. Comput. Phys.* 23, 327–341.
62. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R (1984), . *J. Chem. Phys.* 81, 3684–3690.
63. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. (1997), *Biophys. J.* 72, 1047–1069.
64. van Gunsteren, W. F. Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications; ESCOM: Leiden, The Netherlands: (1989).
65. Straatsma, T. P.; Berendsen, H. J. C.; Stam, A. (1986), *J. Mol. Phys.* 57, 89–95.
66. Helms, V.; Wade, R. C. (1997), *J. Comput. Chem.* 18, 449–462.
67. Hol, W. G.; Halie, L. M.; Sander, C. (1981), *Nature* 294, 532–536.
68. Chothia, C. (1976), *J. Mol. Biol.* 105, 1–14.
69. Hermann, R. B. (1977), *Proc. Natl. Acad. Sci. USA* 74, 4144–4145.
70. Reynolds, J. A.; Gilbert, D. B.; Tanford, C. (1974), *Proc. Natl. Acad. Sci. USA* 71, 2925–2927.
71. Eisenberg, D.; Mclachlan, A. D. (1986), *Nature* 319, 199–203.
72. Lee, B.; Richards, F. M. (1971), *J. Mol. Biol.* 55, 379.
73. Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. (1991), *Science* 252, 106–109.
74. Simonson, T.; Brunger, A. T. (1994), *J. Phys. Chem.* 98, 4683–4694.
75. Hildebrandt, A.; Blossey, R.; Rjasanow, S.; Kohlbacher, O.; P, L. H. (2004), *Phys. Rev. Lett.* 93, 108104.

Chapter 6

Dynamic Protonation Equalibira of Solvated Acetic Acid

(published in *Angew. Chem. Int. Ed.*, **46**, 2939-2943 (2007))

6.1 Summary

For the first time, the dynamic protonation equilibrium between an amino acid side chain analogue and bulk water as well as the diffusion properties of the excess proton were successfully reproduced through unbiased computer simulations. During a 50 ns Q-HOP MD simulation, two different regimes of proton transfer were observed. Extended phases of frequent proton swapping between acetic acid and nearby water were separated by phases where the proton freely diffuses in the simulation box until it is captured again by acetic acid. The pK_a of acetic acid was calculated around 3.0 based on the relative population of protonated and deprotonated states and the diffusion coefficient of excess proton was computed from the average mean squared displacement in the simulation. Both calculated values agree well with the experimental measurements.

6.2 Introduction

The biological functions of many proteins are crucially coupled to protonation equilibria, for instances, in enzymatic reactions such as serine proteases (*1-3*) and carbonic anhydrase (*4*), or in proton pumps in membranes such as bacteriorhodopsin (*5*), cytochrome *c* oxidase (COX) (*6, 7*) and F_OF_1-ATP synthase (*8*). Furthermore, the protein structure itself is often strongly dependent on the predominant protonation states of the titratable side chain groups as well (*9-11*).

In spite of their enormous importance, many aspects of proton transfer (PT) reactions in biomolecules remain poorly understood. Experimental techniques, in particular, are facing fundamental and / or technical difficulties with respect to the direct observation of PT reactions. For example, X-ray crystallography is widely used in determining the three-dimensional structures of biological macromolecules. Nevertheless, hydrogen atoms cannot be detected in most structures except for a few structures at ultrahigh resolution. Although NMR experiments can detect protons

directly, the time resolution of NMR is not short enough to resolve proton transfer processes which occur on time scales as short as tens of femtosecond. Similarly, Neutron diffraction experiments can provide time-averaged proton positions. Mass spectroscopy experiments need to be performed under vacuum conditions and often face the problem of proper peak assignment. Apparently, the only direct experimental observation of PT reactions is from Fourier Transform Infrared (FTIR) Spectroscopy that is able to identify proton transfer paths implicitly when combined with site-directed mutagenesis (*12*). Therefore, it is highly desirable to complement the existing experimental techniques by computational methods.

In the past decades, various computational methods have been developed to calculate the pK_a values of amino acid side chains as well as to perform constant pH simulations of proteins (*13-19*). Hünenberger and co-workers proposed a model using "fractional charges" that allows continuous changes between different protonation states (*13*). Such kinds of fractional models (*13, 14*), however, have also been criticized because of their nonphysical intermediate protonation state (*15, 16*). The work of Brooks and co-workers addressed this problem using a set of continuous titration coordinates that describes transitions between fully protonated or deprotonated states (*17*). Besides the continuous models, several discrete models were proposed by combining Monte Carlo sampling for selecting the protonation states with Poisson-Boltzmann methods (*18, 19*) or thermodynamic integration (*20*) for calculating protonation energies. Recently, Mongan et al. introduced an efficient model that uses the generalized Born (GB) implicit solvation model for the protonation state transition energies and dynamics (*16*). The methods mentioned above allow computing the pK_a of amino acid side chains in a relative efficient manner. However, they do not model explicit proton exchange reactions between the titratable sites and the surrounding aqueous solution or the exchange between different titratable sites. Therefore, these methods may not be suitable to identify proton transfer pathways or to characterize the mechanisms of PT reactions.

This is the area where dynamic simulations of proton transfer come into play. Tuckerman; Marx and co-workers studied the shared proton in hydrogen bonds (*21*) and a hydrated excess proton in water (*22*) using the Car-Parrinello molecular dynamics (CPMD) method (*23, 24*). Lobaugh and Voth investigated proton transport in water by simulating an excess proton in a box of water molecules (*25*) within the centroid molecular dynamics (CMD) (*26*) framework. A similar system was then studied using a multistate empirical valence bond (MS-EVB) model for proton transfer (*27-29*). A recent study presented the dynamic simulation of pK_a values for amino acid side analogues (*30*) using the MS-EVB model and the umbrella sampling technique (*31, 32*). There, different parts of the system phase space were sampled by fixing the distance between the donor and the acceptor (distance between center of excess charge) at different values. The deviation between their computed value and the experimental pK_a was 1-2 pK_a units. Voth and co-workers also studied aqueous proton solvation and transport using the CPMD method (*33*). Besides such model systems, several applications showed the importance and success of studying proton transfer in protein systems by theoretical approaches, for instances, the proton shuttle in green fluorescent protein (*34*), the proton transfer in bacteriorhodopsin (*35*), the proton transfer in Gramicidin A (*36, 37*), the proton transfer along a water chain in the D-pathway of COX (*38*) and the proton translocation in Carbonic Anhydrase (*39*).

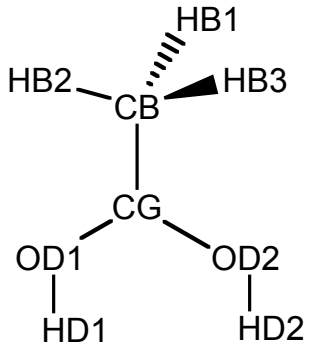
The CPMD simulation mentioned above coupled with the path integral method, as well as the centroid MD technique, provide a quantum description of short-range proton transfer process. However, the transport of a single proton through a biological membrane protein may take as long as 1 ms for systems like bacteriorhodopsin, and the proton has to travel over a distance about 4 to 5 nm across the lipid-bilayer. Such time and length scales, typical for biological systems (especially transmembrane proteins) are currently out of reach for *ab initio* MD approaches. Even though one may resort to biasing techniques to overcome limiting energy barriers, such simulations provide limited insight into the driving forces that activate proton transfer reactions. For example, in the study of Maupin et al. (*30*) the proton transfer pathway and protonation equilibrium between amino acids and aqueous solution and between different solvent molecules could not be directly observed due to the distance constraints employed between the donor and acceptor groups. There is certainly a great need for semi-quantitative approaches that can efficiently explore proton transfer paths in proteins consisting of many subsequent transfer events. Once these transfer paths are identified, more accurate methods can be applied to critically investigate each single step along the pathways.

One such model of intermediate accuracy, the Q-HOP MD method, was introduced earlier to study proton transport in biomolecular systems (*40*-*43*). In the Q-HOP scheme, the proton transfer probabilities for each proton donor and acceptor pair are calculated using a semi-empirical approach during the MD simulation (see the Method section for details). Depending on whether the proton transfer occurs (by comparing the proton transfer probability with a random number), the topology of the system will be modified or be kept unchanged before the next step of MD simulation. This method has been successfully applied to study the proton shuttle in green fluorescent protein (GFP) (*34*) and to understand the mechanism of proton blockage in Aquaporin (*44*). In this manuscript, we present its application to study the explicit protonation equilibrium of solvated acetic acid on a time scale of tens of nanoseconds and at reasonable p*H* condition (p*H* 1). The pK_a of acetic acid is calculated based on the relative population of protonated and deprotonated states observed during the 50 ns Q-HOP MD simulation. The unbiased MD simulations allow to identify the proton hopping mechanism and the driving force of the activated processes of proton transfer.

6.3 Materials and Methods

6.3.1 Parameterization of acetic acid

The segments of protonated and deprotonated acetic acid were constructed based on the segments of protonated and deprotonated aspartic acid in the AMBER force field (*45*). The C_α–C_β bond was deleted and the C_α atom was replaced by an H_β atom. All bonded and non-bonded parameters of the newly added H_β atom were set to the same values as for the other two H_β atoms. The remaining atomic partial charge was added to the C_β atom to make the net charge of the segment integer (0 for protonated acetic acid and –1 for deprotonated acetic acid). The parameterization of acetic acid used in this study is listed in detail in Scheme 1 and Table 1.



Scheme 1: Topology of acetic acid

Atom	Atomic charges in MD			Atomic charges q_{12}^{env} to compute E_{12}^{env}		
	State1[a]	State2	State3	State1	State2	State3
HB (1-3)	0.0488	0.0488	-0.0122	0.0488	0.0488	-0.0122
CB	-0.0743	-0.0743	-0.1600	-0.0743	-0.0743	-0.1600
CG	0.6462	0.6462	0.7994	0.9462	0.9462	0.7994
OD1	-0.5554	-0.6376	-0.8014	-0.8554	-1.2376	-0.8014
HD1	dummy	0.4747	dummy	dummy	1.0747	dummy
OD2	-0.6376	-0.5554	-0.8014	-1.2376	-0.8554	-0.8014
HD2	0.4747	dummy	dummy	1.0747	dummy	dummy

Table 1 Atomic charges for acetic acid. [a]State1: protonated on OD2; State2: protonated on OD1; State3: deprotonated.

6.3.2 Q-HOP MD simulation for solvated acetic acid

In our Q-HOP method, the proton hopping probability p is calculated from the energy difference E_{12} between the pair of protonated donor/deprotonated acceptor and the pair of deprotonated donor/protonated acceptor, and from the distance between donor and acceptor atoms, R_{DA} (*42*), see equations (4)-(6) below. Depending on the values

of E_{12} and R_{DA}, two different approaches are used to compute the hopping probability (transfer rate) (*40*).

For large E_{12} and large R_{DA}, a modified transition state theory is used accounting for the zero-point energy and the tunneling effect:

$$p = \kappa(T, E_M)\frac{k_B T}{h}\exp\left(-\frac{E_b - h\omega/2}{k_B T}\right)\Delta t \qquad (1)$$

where $\kappa(T, E_M)$ is the enhancement of the classical transfer rate due to tunneling $\kappa = k_{QM}/k_{classical}$ as function of temperature T and E_M (*40*). $E_M = E_{\max} - \max(E_{\min,1} - E_{\min,2})$ is the difference between the energy maximum $E_{\max}$ and the larger one of the two energy minima $E_{\min,1}$ and $E_{\min,2}$ along the two-well potential of a typical transition reaction. $h\omega/2$ is the zero-point energy obtained by considering the bonds that contain a transferring proton as a quantum-mechanical harmonic oscillator with frequency ω at the educt well minimum (*42*). E_b is the energy barrier along the two well potential and is calculated in Q-HOP as a function of E_{12} and R_{DA}:

$$E_b(E_{12}, R_{DA}) = S(R_{DA}) + T(R_{DA})E_{12} + V(R_{DA})E_{12}^{\ 2} \qquad (2)$$

where *S*, *T* and *V* have a simple functional dependence on R_{DA} (*42*).

We note that this high-barrier regime is only a limiting case. In nanosecond time-scale simulations, the probabilities computed from eq (1) are too low for PT events to occur. For barriers involving small E_{12} and small R_{DA}, the transfer rate is estimated by following the propagation of a one-dimensional wave package as a solution of the time-dependent Schrödinger equation and computing the fractional population that crosses the barrier:

$$p = 0.5\tanh(-K(R_{DA}))E_{12} + M(R_{DA}) + 0.5 \qquad (3)$$

where *K* and *M* are also functions of R_{DA} (*40*).

In the Q-HOP method, E_{12} is a sum of two contributions:

$$E_{12} = E_{12}^0 + E_{12}^{env} \qquad (4)$$

E_{12}^0 is the energy difference between a donor-acceptor pair in vacuum. It is obtained from the following empirical relationship (*42*),

$$E_{12}^0 = \alpha + \beta \cdot R_{DA} + \gamma \cdot R_{DA}^2 \qquad (5)$$

where α, β and γ are fitted parameters compiled in a recent data set involving MP2/6-31++G** calculations of all titratable amino acids (*46*). The values for Asp-H_2O are 277.2 kcal/mol, -171.4 kcal/(mol·Å) and 19.9 kcal/(mol·Å^2) for α, β and γ, respectively. Very similar results were obtained for acetic acid in *trans* and *cis* conformations. The deviations were less than 2.3 kcal/mol. Therefore, only a single

parameter set, that of *trans* acetic acid, was used. The environmental contribution E_{12}^{env} is calculated from the coulombic interactions between the two pairs and the environment:

$$E_{12}^{env} = E_{DH-A,env}^{Coul} + E_{D-AH,env}^{Coul} \tag{6}$$

Here $E_{DH-A,env}^{Coul}$ and $E_{D-AH,env}^{Coul}$ are defined as:

$$E^{Coul} = \sum_{i}^{donor\_acceptor\_atoms} \sum_{j}^{remaining\_system} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} \tag{7}$$

q_i and q_j are the respective atomic partial charges, r_{ij} is the atomic distance between atoms i and j, and ε_0 is the permittivity of vacuum. When using the atomic partial charges from the AMBER force field, E_{12} calculated by equation (4)-(6) showed systematic deviations from reference QM/MM calculations (see below) of 8 to 25 kcal/mol. Therefore, we decided to optimize a separate set of q_{12}^{env} charges to reproduce the results from the QM/MM calculation. A similar refinement was recently introduced in the MS-EVB model as well (*30*). We stress that these charges are not used in the propagation of the trajectory (normal MD part), but only serve to compute the environmental contribution from equation (6).

6.3.3 Generation of favorable transfer geometries

To generate favorable transfer geometries for the charge fitting (see below), two molecular dynamics (MD) simulations were performed. One pair each of a protonated acetic acid and a water molecule (denoted by AcAH-H2O) and of a deprotonated acetic acid and a hydronium ion (denoted by AcA^--$H3O^+$) were solvated in cubic boxes of 24 Å box dimension, using SPC/E water model (*47*). During the subsequent MD simulations at 300K, NPT conditions, the atomic distance between the O_δ atom of acetic acid and the OW atom of water molecule/hydronium ion was restrained to 2.5 Å. All other settings of the two simulations were the same as in the Q-HOP simulations. Each simulation was conducted over 50 ps and 50 snapshots were selected from each simulation for the charge fitting.

6.3.4 Quantum mechanics/molecular mechanics (QM/MM) calculation for charge fitting

QM/MM calculations were performed on these 100 snapshots to compute the energy differences between the AcAH-H2O pair and the AcA^--$H3O^+$ pair in the same geometry (E_{12}). All QM/MM calculations were performed using the NWChem 4.7 package (*48, 49*). Each pair and its surrounding water molecules within 6.7 Å were treated at density functional theory level using the B3LYP/ 6-31++G** functional/basis set. The remaining water molecules in the system were modeled by point charges taken from the AMBER99 force field (*45*). 10 geometries (5 from each simulation) were also computed at MP2/6-31++G** level and a systematic difference

was found: the AcAH-H2O pair was more favorable at the MP2 level by about 1.5 kcal/mol. The charge fitting calculations were performed using our Q-HOP method (see below) to reproduce the E_{12} values from the QM/MM calculation and accounted for the systematic difference between DFT and MP2 calculations. It was found that the fitted charges are not very sensitive to the exact value of E_{12}. Therefore, we only needed to characterize E_{12} to a reasonable scale. The details of the charge fitting will be reported in a separate manuscript. The atomic charges employed for computing E_{12}^{env} are listed in the supplementary material. Note that only the atomic charges of C_{γ}, $O_{\delta 1,\,2}$ and $H_{\delta 1,\,2}$ of acetic acid were modified. The charges on H_2O and H_3O^+ were left unchanged.

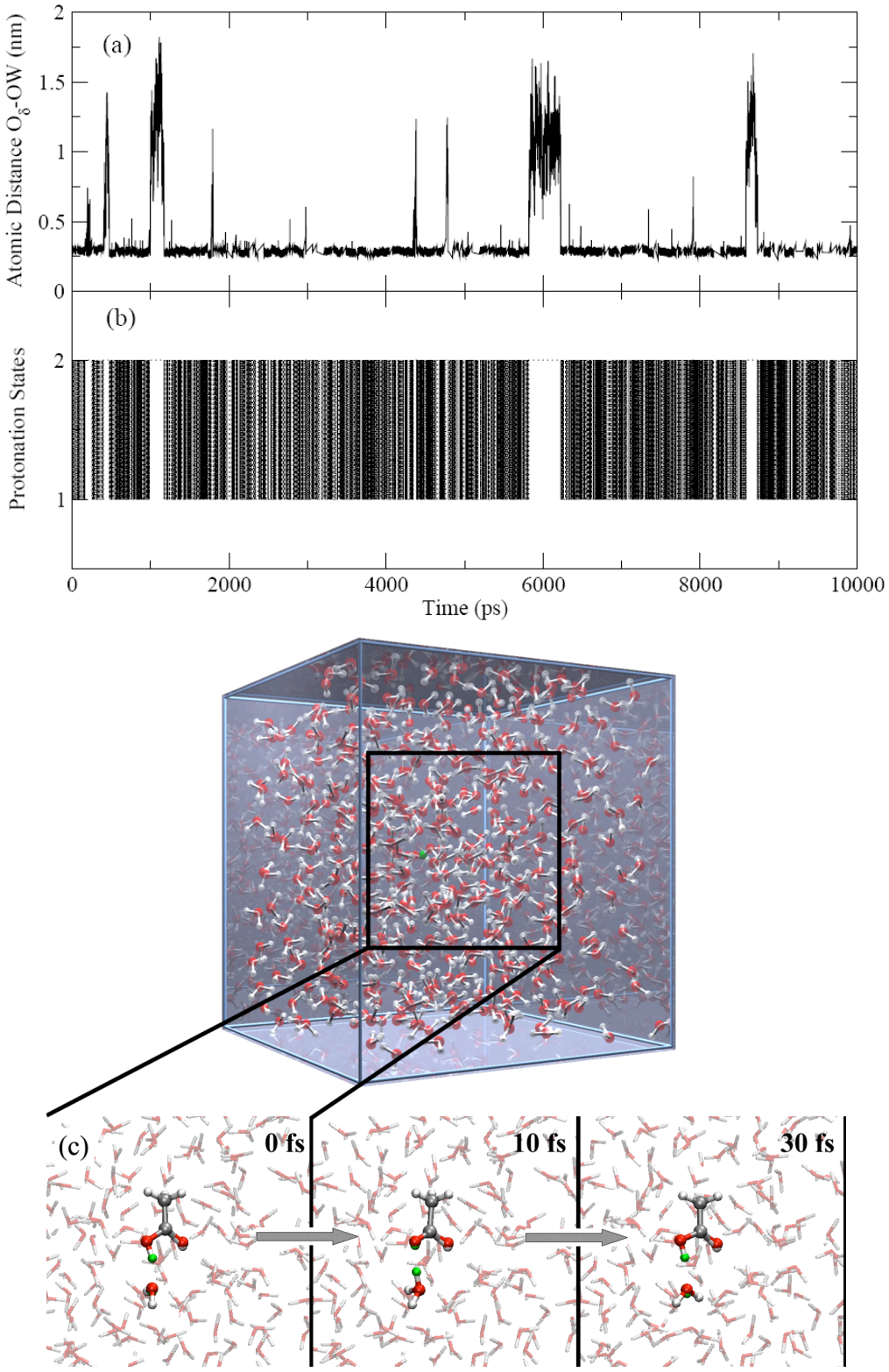
The Q-HOP MD simulations as well as the single point calculations for charge fitting were performed using a modified version of the NWChem 4.7 package employing the AMBER99 force field (*45*). In this implementation, the Particle Mesh Ewald (PME) method (*50*) is used for calculating long-range electrostatic interactions during molecular dynamics. For evaluating E_{12}^{env}, all coulombic interactions were computed between the donor–acceptor pairs and the other atoms of the simulation box. In the Q-HOP MD simulations, the AcAH-H2O pair and the AcA^--$H3O^+$ pair (same as in the MD simulation for generating favorable hopping geometries) each were solvated in cubic boxes of 24 Å side length, using SPC/E water molecules (*47*). All coordinate sets were first subjected to 500 steps of steepest-descent energy minimization (Q-HOP switched off). The solvent and modeled residues were then relaxed during a 100 ps MD simulation at 300 K prior to the Q-HOP MD simulation. Then two 10 ns Q-HOP MD simulations were performed on each system. After observing that both simulations gave very similar results, one simulation with AcA^--$H3O^+$ as the starting pair was extended to 50 ns. All analyses shown here are based on this 50 ns simulation. During the simulation, temperature (300 K) and pressure (1 atm) were maintained by weak coupling to an external bath (*51*). The SHAKE procedure (*52*) was applied to constrain all bonds that contain hydrogen atoms. Non-bonded interactions were treated using a cutoff of 10 Å and long-range electrostatic interactions were computed using the PME method as mentioned before. The time step of the simulations was 1 fs throughout. Scanning for possible proton transfer events was performed every 10 steps and snapshots were also recorded every 10 steps to track all hopping events. All water molecules as well as the acetic acid were possible donor/acceptors. All protons of the water molecules were transferable.

6.4 Results

6.4.1 Protonation equilibrium between acetic acid and the water molecules of a water box

During the Q-HOP MD simulation, two different protonation equilibria were observed. Figure 1a and Figure 1b show the position of the “free” proton and the distance between the hydronium ion and the deprotonated acetic acid (when the proton stays on hydronium ion). Figure 1a and 1b only show the first 10 ns for better visibility. The results for the full length of the simulation are presented in Figure 1e.

Two different situations can be distinguished. The first type of protonation equilibrium “proton swapping” only involves the acetic acid and a nearby water molecule/hydronium ion forming a hydrogen bond with acetic acid. The other type of protonation equilibrium “traveling proton” is that of an excess proton and all water molecules of the simulation box (see Figure 1c and Figure 1d).



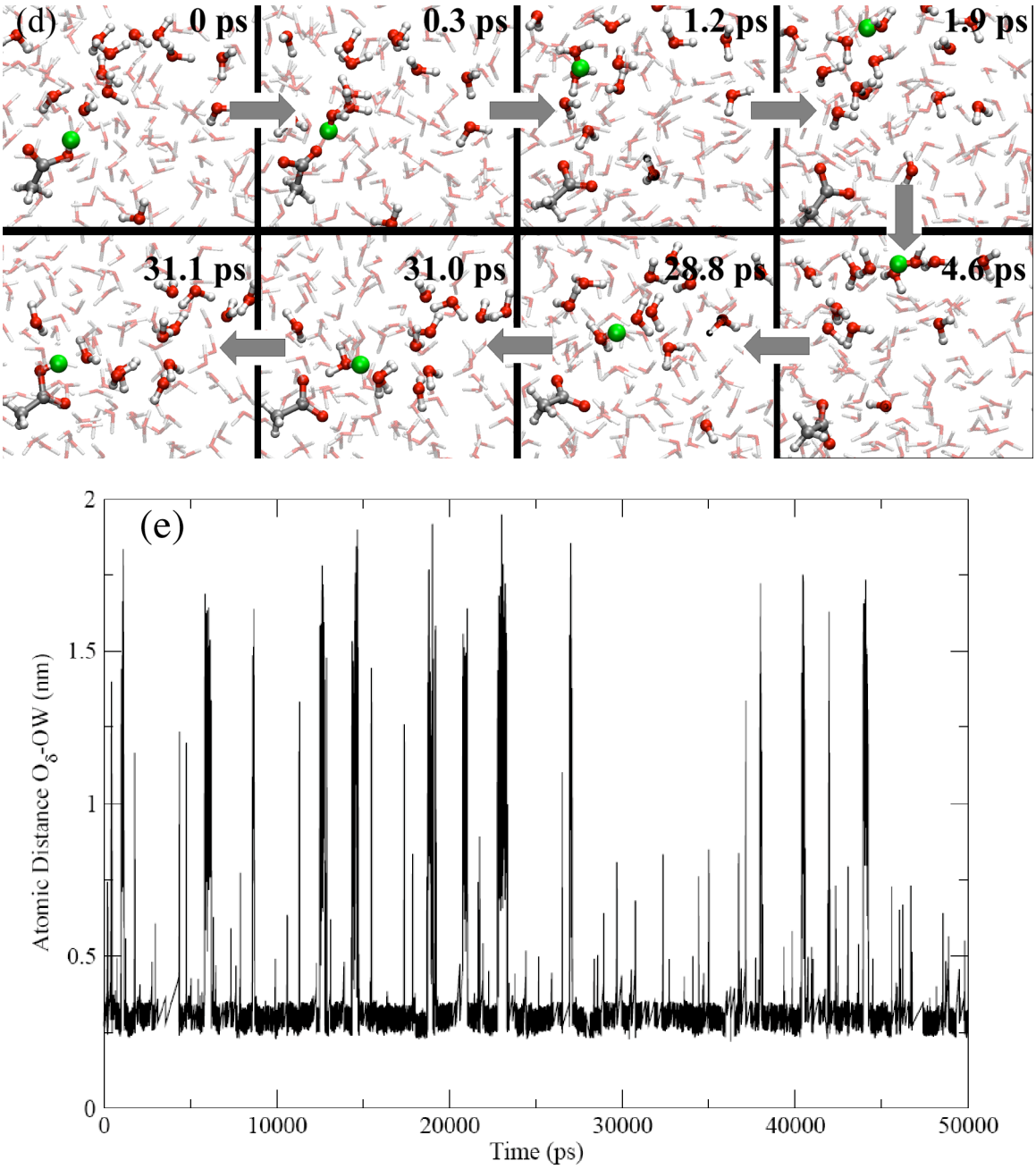


Figure 1: (a) Time evolution of atomic distance between the O_δ atom of acetic acid and the OW atom of the hydronium ion (of the first 10 ns); (b) Time evolution of the protonation states of the system: 1 denotes the state in which the proton is located on the acetic acid, 2 denotes the state in which the proton is bound to the hydronium ion; (c) Snapshots of the "fast-swapping" phase (the simulation box is shown at the top and the donor/acceptor pair is enlarged); (d) Snapshots of a typical "traveling" phase lasting 31 ps involving 16 different water molecules. Only the first 4 and the last 3 transfer steps and the molecules involved are shown. In (c) and (d), oxygen atoms are colored in red and protons that were transferred are colored in green. (e) Time evolution of atomic distance between the O_δ atom of acetic acid and the OW atom of the hydronium ion (full length).

The first scenario accounts for more than 90% of the total proton transfer events

(Figure 1b). Whereas the proton is energetically more favorable on acetic acid, it may frequently "visit" the bound water molecule driven by environmental fluctuations (see below). In most cases, it will almost immediately hop back to acetic acid. During this fast proton swapping, the population of protonated acetic acid is much higher (97%) than the population of hydronium ions. This reflects the fact that protonated acetic acid is energetically more favorable than hydronium ion.

In some cases of the fast proton swapping, the proton did not hop back to acetic acid, but escaped to another water molecule hydrogen-bonded to the hydronium ion. This process may continue, and the proton starts "traveling" in the water box (see the peaks in Figure 1a). Such traveling periods lasted from a few picoseconds to hundreds of picoseconds before the proton eventually hops back to the acetic acid. They were observed several times every nanosecond. The total time spent with traveling amounts to slightly less than 7% of the total simulation time. Although we did not test this sofar, it seems very plausible that the duration of the traveling events depends on the size of the simulation box in a proportional manner. Figure 1c and 1d illustrate characteristic snapshots of both scenarios.

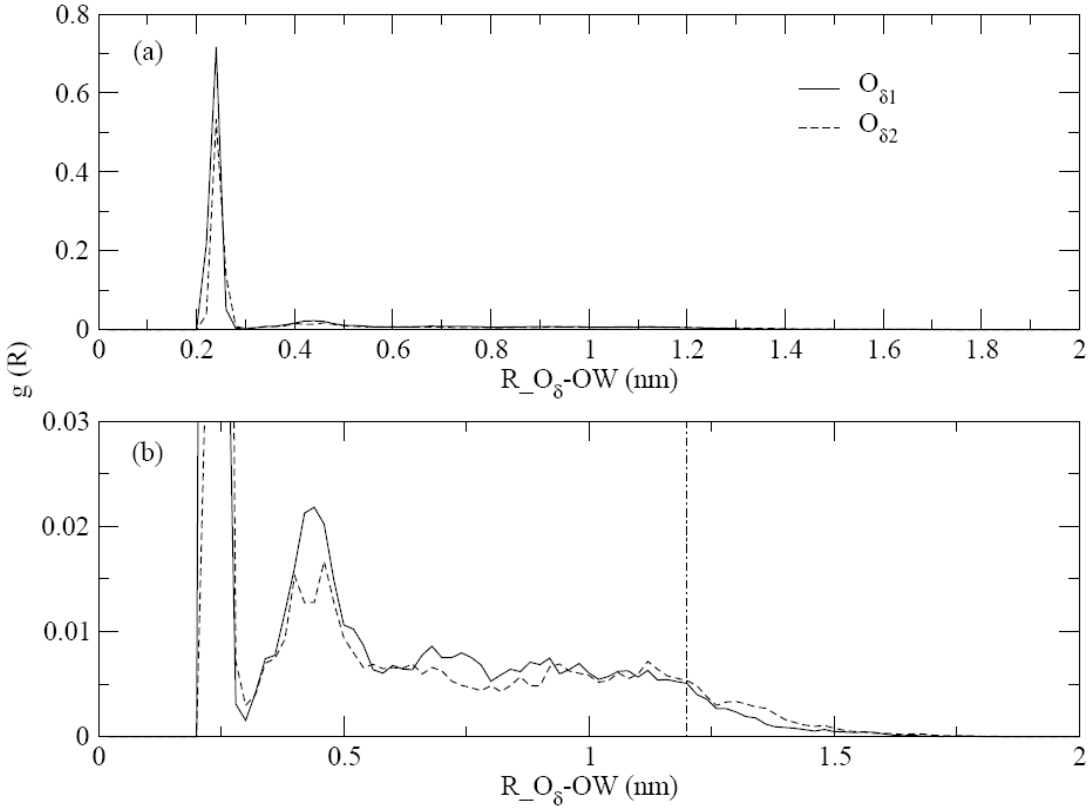


Figure 2: (a) Radial distribution (normalized) of hydronium ions around the two carboxyl oxygen atoms of the acetic acid; (b) Amplification of (a).

Figure 2 shows the radial distribution of hydronium ions around the two carboxyl oxygen atoms of the acetic acid. The first dominant peak at 2.4–2.6 Å is due to the fast proton swapping between the acetic acid and bound water molecules mentioned above. A small second peak appears at 4–5 Å (see Figure 2b), which belongs to the first solvation shell of the AcA^--$H3O^+$ pair. The distribution function then becomes

flat between 5 Å to 12 Å, and drops slowly to 0 for even longer distance. The flat distribution indicates that the hydronium ion is uniformly distributed when the proton is traveling in the simulation box. This uniform distribution allows us to estimate the traveling time of the proton for different box sizes. When the box size is much larger than 5 Å (where the second peak ends), the average traveling time should be proportional to the volume of the box. The drop of the distribution function beyond 12 Å results from the corner effects of the cubic box with 24 Å dimensions. The dashed and solid lines show the radial distributions separately computed for both carboxylic oxygen atoms of acetic acid. The difference between both lines gives an indication of the statistical error of the simulation.

6.4.2 Proton hopping events

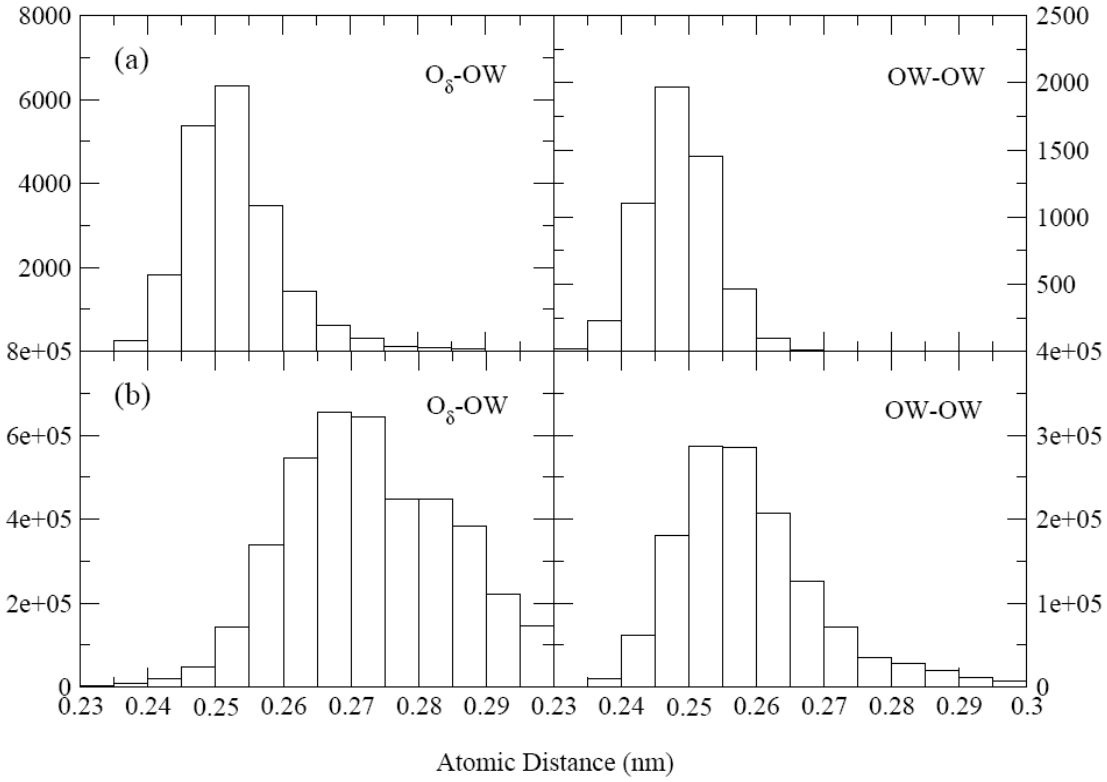


Figure 3: Atomic distances between the O_δ atom of acetic acid and the OW atom of water/hydronium ion and between OW atoms of hydronium ion and the closest water molecule. (a) in hopping events (b) in non-hopping events.

Figure 3 shows the binned distributions of the atomic distance between the O_δ atom of acetic acid and the OW atom of water molecule/hydronium ion as well as the atomic distance between the OW atom of hydronium ion and the OW atom of its closest water molecule in non-hopping (a, c) and hopping events (b, d). In the standard case (non-hopping), the O_δ–OW distance ranges from 2.5 Å to 3.0 Å and shows a peak at 2.7 Å; the OW–OW distance ranges from 2.4 Å to 2.9 Å and shows a peak at 2.5–2.6 Å. Both distances decrease by about 0.1 to 0.15 Å in the hopping events (see Figure 3b). This indicates that proton transfer occurs at relatively closer distances in both

acetic acid–hydronium/water pair and hydronium–water pair. This agrees well with intuition and with our previous studies of proton transfer processes on model systems using electronic structure methods, where we found that the energy barrier decreases when donor and acceptor approach each other (*42*).

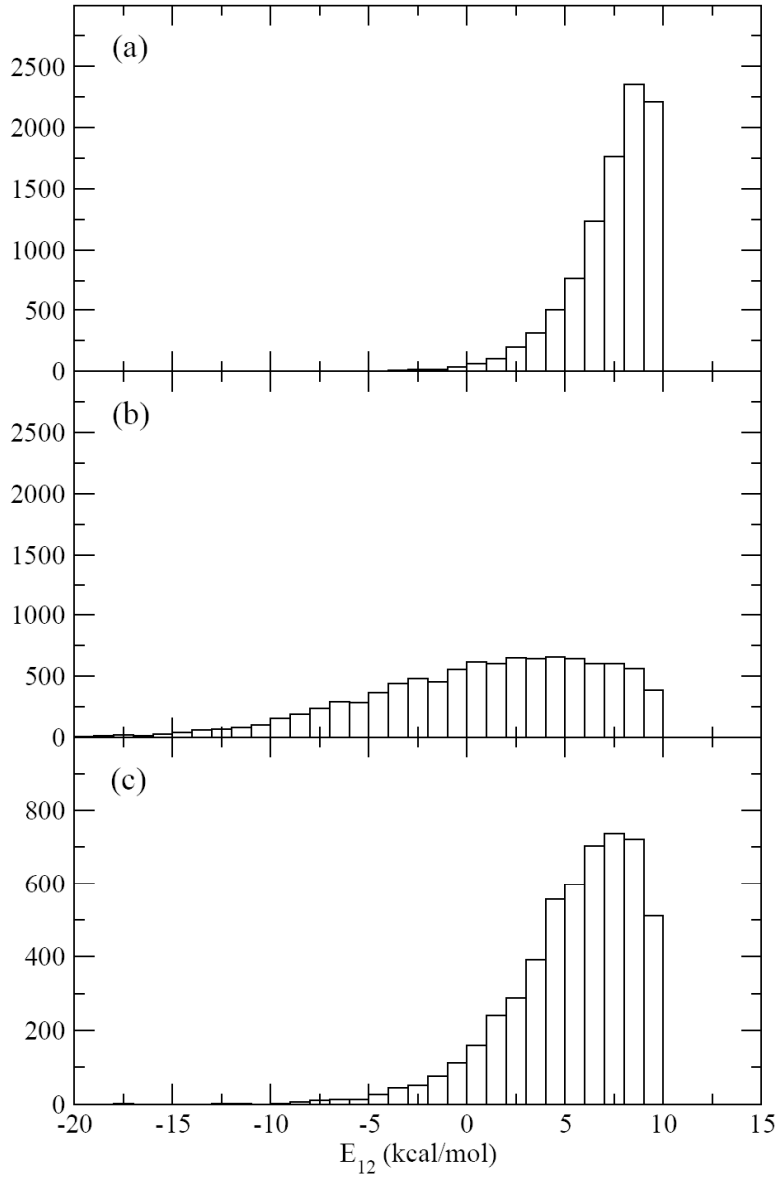


Figure 4: Distribution of E_{12} during the simulation. (a) hopping events when proton hops from acetic acid to water; (b) hopping events when proton hops from water to acetic acid; (c) hopping events between different water molecules.

The energy difference between the protonated donor_deprotonated acceptor pair and the deprotonated donor_protonated acceptor pair, E_{12} is shown in Figure 4. When the proton hops from the acetic acid to a water molecule, E_{12} ranges from 0 to 10 kcal/mol and the average value is about 7.5 kcal/mol (Figure 4a). For the reverse case (from hydronium ion to acetic acid), E_{12} shows a broad range from –20 kcal/mol to 10 cal/mol (Figure 4b). For the case of proton swapping between different water molecules, E_{12} ranges from –5 kcal/mol to 10 kcal/mol (see Figure 4c). These results

illustrate that, for the system we studied, proton hopping mostly took place when E_{12} is less than 10 kcal/mol. These findings also agree with our previous QM calculations on model systems (*40*). The generally positive E_{12} value in the upper panel indicates that protonated acetic acid is energetically more favorable than a solvated hydronium ion. The proton can only be transferred to the hydronium ion when the energy difference becomes small enough. For the reverse process, on the other hand, the broad range of E_{12} values, especially for many cases in which E_{12} is less than –5 kcal/mol, indicates that this process takes place not only because the energy differences are small, but also because the protonated acetic acid is more favorable. In contrast, there are no energetic preferences for the proton transfers between the hydronium ion and water molecules, since both donor and acceptor are obviously modeled by the same set of parameters.

6.4.3 Proton hopping and hydrogen-bonding network

Figure 5 shows a typical scenario of proton transfer between acetic acid and hydronium/water. To understand the mechanism of the proton hopping, we examined the hydrogen-bonding network of a subsystem (see Figure 5). This subsystem consists of the donor–acceptor pair and three water molecules that form hydrogen bonds with either the donor or the acceptor atom. Here, Wat1 and Wat2 denote the two closest water molecules that form hydrogen bonds (as acceptors) with the hydronium ion and Wat3 denotes the closest water molecule that forms a hydrogen bond (as donor) with the O_δ of acetic acid. Figure 6a displays the average distances between the hydrogen-bonding atoms in different protonation states as well as scenarios right before the proton hopping. The three atomic distances between OW of Wat1–Wat3 and their hydrogen-bonding partners are labeled as d1, d2 and d3. R_{DA} denotes the atomic distance between the donor and acceptor atoms. We start the discussion from the left side in a situation in which the proton resides on the hydronium ion, and the three water molecules Wat1 – Wat3 are located very close to their hydrogen-bonding partners (d1: 2.54±0.09 Å; d2: 2.67±0.14 Å; d3: 2.83±0.22 Å). Also the donor and acceptor atoms are at very close distance (R_{DA} 2.50±0.09 Å). These close contacts stabilize the charge-separated state of the system. At the time intervals 10 fs before and at the time of proton transfer to the acetic acid, all three water molecules are displaced from the donor–acceptor pair (d1: 2.62±0.12 Å, 2.59±0.11 Å; d2: 2.80±0.17 Å, 2.78±0.17 Å; d3: 3.02±0.24 Å, 3.00±0.25 Å). At this stage, also the donor and the acceptor are found at slightly larger distance (R_{DA} 2.53±0.13 Å, 2.55±0.08 Å). This concerted movement destabilizes the actual protonation state and the proton hopping is facilitated. After the proton is transferred to the acetic acid, the environment quickly adapts to the new protonation state (d1: 2.73±0.12 Å; d2: 2.89±0.16 Å; d3: 3.04±0.24 Å). One pronounced change is the significantly enlarged distance between donor and acceptor, R_{DA} that increases from 2.55±0.08 Å to 2.75±0.14 Å. Considering the back-transfer from acetic acid to water, the main driving force seems again a strong decrease of the R_{DA} distance accompanied by small decreases of the water distances.

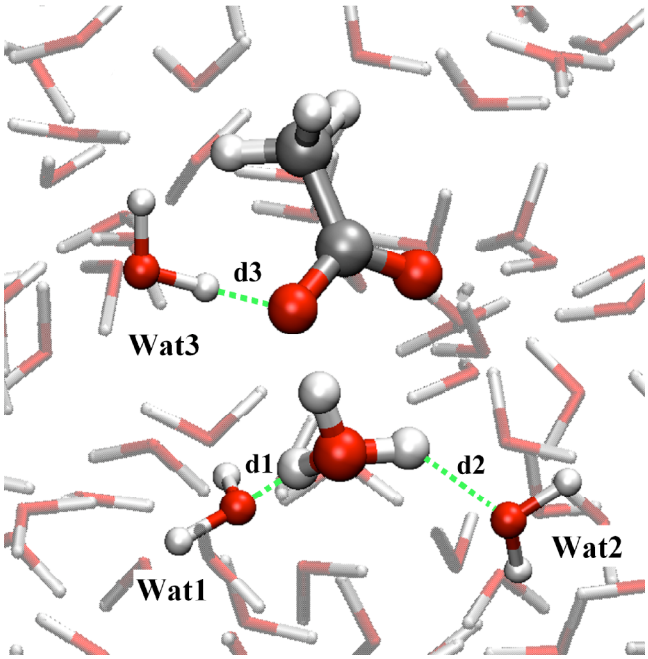
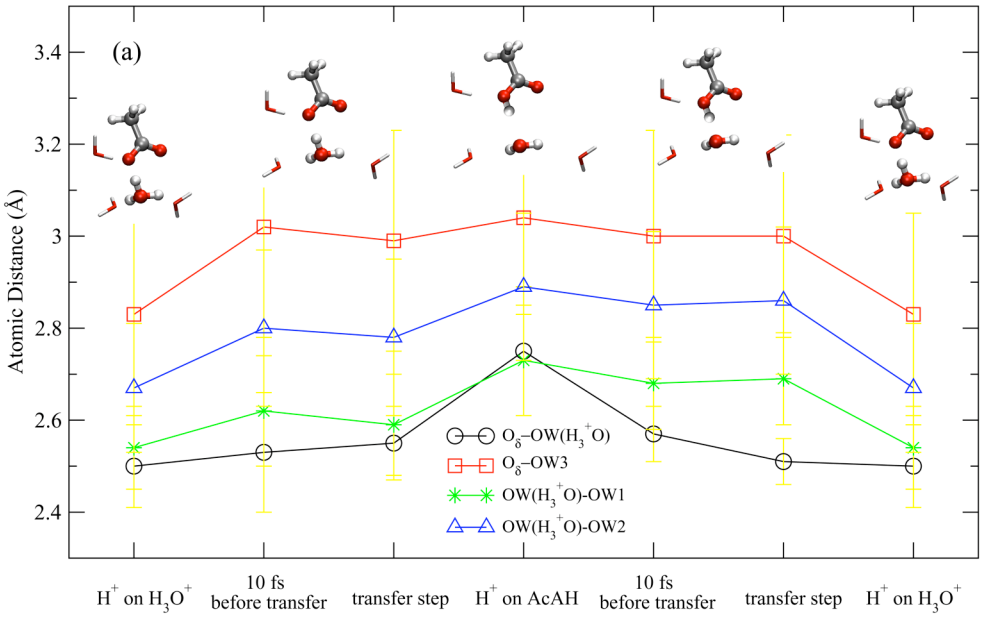


Figure 5: A snapshot of a typical transfer scenario observed during the simulation. The donor (hydronium ion) and the acceptor (acetic acid) as well as the three closest water molecules are shown using "atom and bond" form. Oxygen atoms are colored in red and protons are colored in white. The green dash lines represent possible hydrogen bonds.



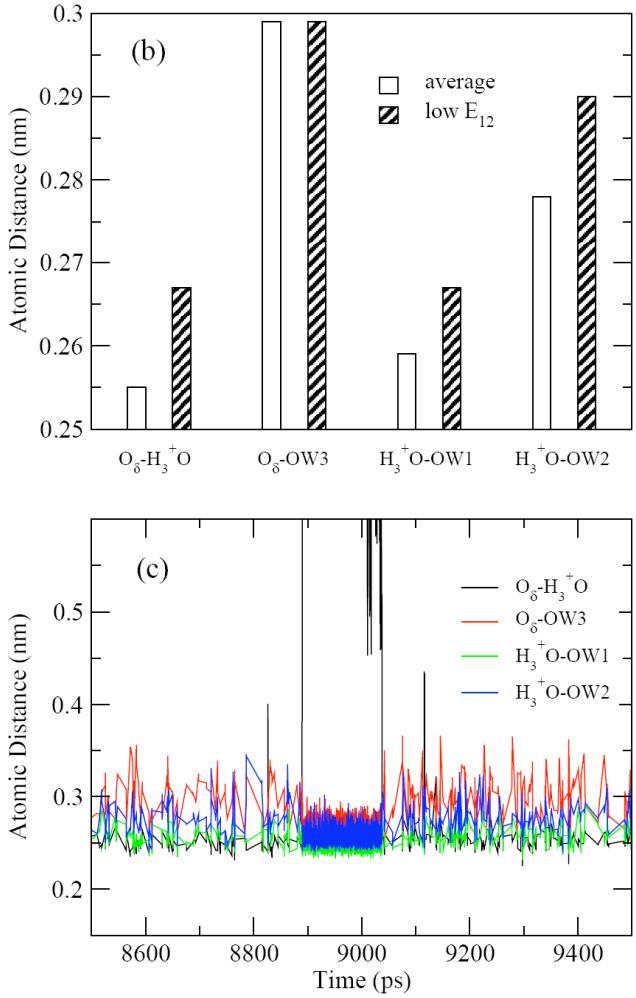


Figure 6: (a) Evolution of the atomic distances R_{DA}, d1, d2 and d3 (see text for definition) (a) in different transfer steps; (b) in low E_{12} hopping events; (c) in different protonation equilibria ("fast swapping" and "traveling").

We also analyzed the dynamic evolution of d1, d2 and d3 in the two types of protonation equilibria — in the fast swapping case and in the proton traveling case. An example is shown in Figure 6b: In the fast swapping case, all three distances show large fluctuations. As soon as the proton starts traveling between different water molecules, d1–d3 become stable and only fluctuate slightly around their average value. Figure 6c shows the comparison between the low E_{12} hopping events (E_{12}<-10 kcal/mol) and the average of all hopping events. It is obvious that the low E_{12} hopping events took place when the hydrogen-bonding network around the hydronium ion was strongly weakened.

6.4.4 Environmental effects and activated processes

When considering forth-and-back proton swapping as a cyclic process:

$$AcAH + H_2O \xrightarrow{\vec{E}_{12}} AcA^- + H_3O^+ \quad \xrightarrow{\tau} \quad AcA^- + H_3O^+ \xrightarrow{\overleftarrow{E}_{12}} AcAH + H_2O$$

we can exam the sum of $\vec{E}_{12}$ and $\overleftarrow{E}_{12}$ as a function of the lag time τ between the two

transfer events. As shown in Figure 7a, most of the proton swapping cycles (>95%) finished in relatively short time (τ<100 fs). Negative values of the sum are mostly found for short τ events. This indicates that during short cycles, the environment sometimes did not relax to the altered protonation state. However, for longer lag times (τ>100 fs), we find in most cases considerable relaxation of the environment. The asymmetric distribution of positive and negative sums for short lag times illustrates that the relaxation of the environment may occur during very short times after the proton hopping occurred.

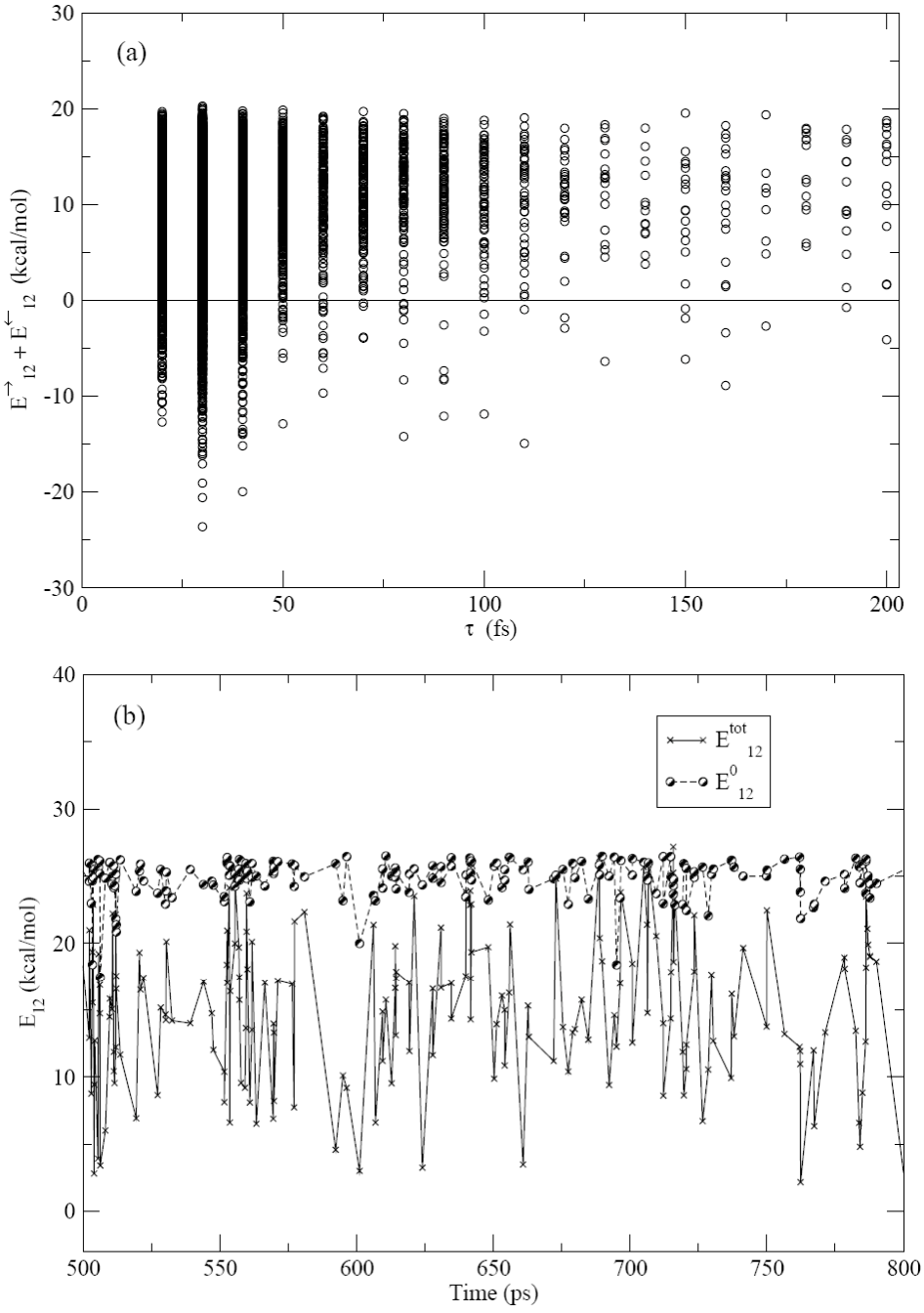


Figure 7: (a) Sum of $\vec{E}_{12}$ and $\overleftarrow{E}_{12}$ as a function of the lag time τ between the two forth-and-back proton swapping events; (b) Environmental contributions to the total E_{12} by comparing E_{12} and E^{0}_{12}.

As shown in Figure 3 and Figure 6a, the proton transfer between acetic acid and hydronium ion/water occurs mostly when the donor acceptor distance R_{DA} is less than 2.6 Å. However, the short R_{DA} cases only occupy a small fraction of the whole simulation (about 12%), which indicates an activated process. In addition, short R_{DA} alone cannot evoke the activated process (see Figure 6a). The surrounding environment also has to be ready to stabilize the new protonation state. Figure 7b shows an example of the proton transfer from acetic acid to water molecule. Clearly, the total E_{12} is dramatically lower than E_{12}^{0} due to addition of the environmental contributions. A fully solvated environment is therefore crucial for characterizing the protonation equilibrium even for a small molecule like acetic acid.

6.4.5 Estimating pK_a from the relative population of protonated acetic acid

As mentioned before, acetic acid was protonated during about 90% of the total simulation. In the current simulation setup, the concentration of free proton (hydronium ion) is 0.1 mol/L when the proton stays on the hydronium ion and 0 when it stays on the acetic acid. Following the classical definition of pK_a,

$$pK_a = -\log\frac{[H^+][A^-]}{[HA]} \tag{8}$$

the pK_a value of acetic acid estimated from the observed populations in our simulation is about 3.0. By dividing the entire trajectory into 25 windows (2 ns each, see Table 2 for details), we derive an average relative population of AcAH of 0.90 ± 0.09 (see Table 2). This corresponds to pK_a values from 2.4 to 4.5. The experimental pK_a of aspartic acid side chain ranges from 3.6 to 4.5 and the experimental pK_a of acetic acid is 4.7. Since we didn't fit any simulation parameters to reproduce the experimental pK_a value, we consider the computed pK_a a very acceptable value. Certainly, other computational techniques allow computing the average pK_a much more precisely at a much lower computational cost (*16, 18, 19*). However, besides the pK_a value, the Q-HOP concept also provides insights into time-dependent processes and allows identifying the driving forces of the activated processes of proton transfer. Moreover, one can obtain an idea of possible mechanistic proton transfer pathways. If desired, individual cases along the discovered pathways can then be studied at much higher accuracy using electronic structure theory coupled to rate theories such as variational transition state theory (*53*).

6.4.6 Diffusion coefficient of the excess proton

Another possible comparison between our simulation and experimental observation is the diffusion coefficient of the excess proton. The trajectory containing the traveling hydronium ion (not the fast swapping case) was divided into 11 parts (large windows) that are each 200 ps long. Then a sliding window of 10 ps was used to calculate the displacement of the excess proton for each large window. The diffusion constant was then computed using the Einstein relation,

Window	Life time of H_3O^+ (ps)	Relative population of AcAH[a]
1	405.33	0.80
2	28.05	0.99
3	267.24	0.87
4	256.78	0.87
5	169.20	0.92
6	86.09	0.96
7	441.86	0.78
8	506.79	0.75
9	65.43	0.97
10	467.37	0.77
11	336.44	0.83
12	591.37	0.70
13	22.64	0.99
14	185.87	0.91
15	54.20	0.97
16	43.13	0.98
17	46.10	0.98
18	109.52	0.95
19	163.83	0.92
20	110.27	0.95
21	312.64	0.84
22	187.33	0.91
23	284.26	0.86
24	27.28	0.99
25	46.74	0.98

Table 2 Relative population of protonated acetic acid in different simulation windows. [a]Relative population of AcAH = (window size - Life time of H_3O^+) / window size

$$D = \left\langle \frac{| r(t) - r(0) |^2}{6t} \right\rangle \tag{9}$$

where the averaging was performed within a large window. Our calculated result ($9.29 \cdot 10^{-1} \pm 1.43 \cdot 10^{-1}$ Å^2 ps^{-1}) nicely agrees with the experimental value of $9.3 \cdot 10^{-1}$ Å^2 ps^{-1}. By choosing different sizes of large window and sliding window, we obtain slightly different results (see Figure 8). Diffusion constants calculated using different window sizes mostly fall in the range of $9.0 \cdot 10^{-1}$ Å^2 ps^{-1} to $9.5 \cdot 10^{-1}$ Å^2 ps^{-1}, which shows that the value is essentially independent of the size of the windows. This agreement indicates that the Q-HOP MD simulation correctly characterizes the dynamics of the excess proton.

6.5 Discussion

The main objective of this study was to show that current simulation methodology allows simulating the dynamic protonation equilibrium between an amino acid side chain analogue and a bulk water surrounding. In contrast to a previous study by Voth and colleagues (*30*), which employed the Potential of Mean Force (PMF) technique, it was not necessary to impose a certain proton transfer pathway in our simulation. Instead, the pathway and the driving forces will be revealed by the simulations. Therefore, one can easily extend the current simulation methodology to real protein systems, where the proton has to travel over large distances and the proton transfer pathway is not known beforehand.

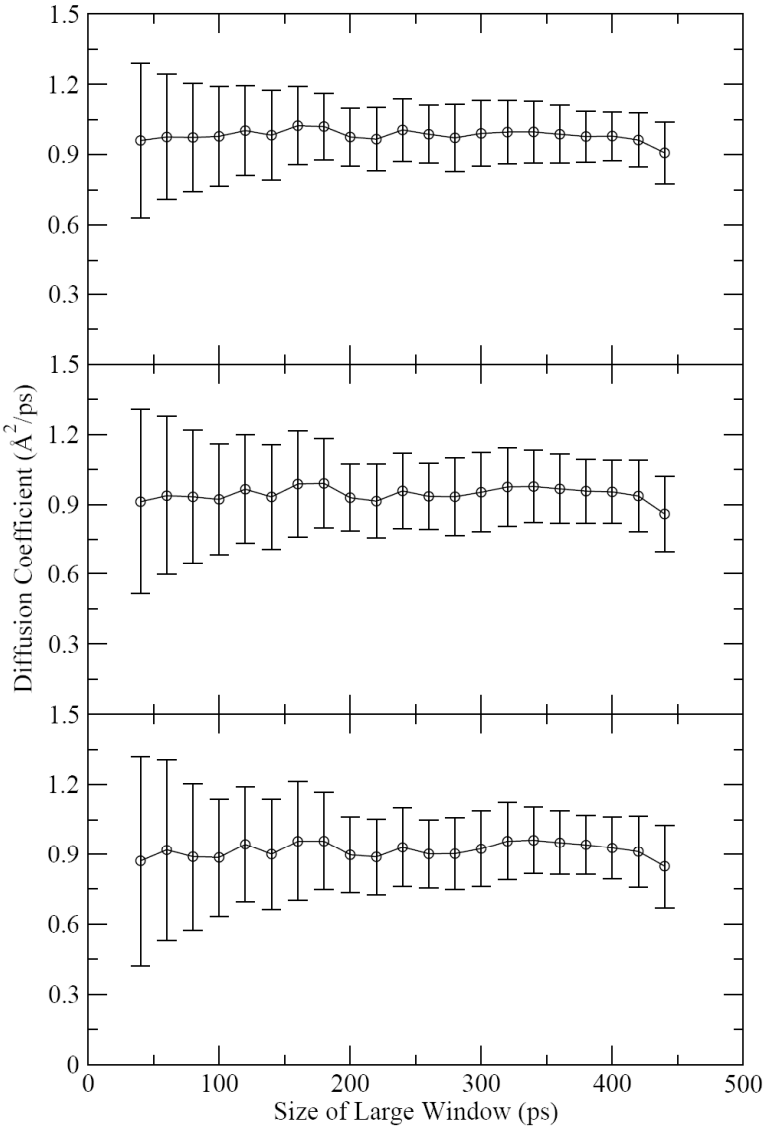


Figure 8: Diffusion coefficient of the excess proton calculated using different large window sizes and different sliding window sizes. From top to bottom, the sliding window sizes are 5 ps, 10 ps and 15 ps, respectively.

6.5.1 Sufficient sampling and time-scale of the simulation

It is always a crucial point for simulation studies to ensure whether the sampling of possible states of the simulated system is enough to characterize the properties of the system or not. As shown in the analysis of our simulation, e.g. the radial distribution of hydronium ions as well as the pK_a calculation, the reported properties were sufficiently sampled during the simulation. In Figure 2, the slightly different distribution curves for both oxygen atoms of the carboxyl group of the acetic acid give an indication of the statistical errors of the sampling. On the other hand, the flat proton distribution between 5 Å to 12 Å shows that the sampling is adequate. Another test is the stability of the calculated pK_a value of acetic acid. By dividing the simulation into two-ns-windows, a statistical error of 9% for the relative population of protonated acetic acid was obtained.

As mentioned above, the protonation equilibria observed in the simulations can simply be interpreted as periods of the proton being confined within a small region around the acetic acid, and periods of traveling. Using the small simulation box of this work, traveling periods can last over hundreds of picoseconds. This is closely related to the bulk diffusion coefficient of the excess proton in water (*41*). The uniform radial distribution of hydronium ions around the acetic acid allows us to estimate the traveling time of the proton for different box sizes, as the average traveling time should be proportional to the volume of the box for reasonable large box. As found in our simulations, it took tens of nanoseconds to observe traveling events at various lengths. Since the computational cost of the Q-HOP method is comparable to classical MD simulations, simulations on this time scale are feasible for many applications. Whereas other more accurate methods, e.g. QM/MM methods (*54*) as well as *ab initio* MD (*21, 22*), are still too expensive for these kinds of studies.

6.5.2 Limits of the Q-HOP method

Currently, the Q-HOP method is not able to handle the case of a proton shared between donor and acceptor molecules. In our simulations, such proton sharing shows up as frequent exchange of proton between donor and acceptor. However we do not expect that this frequent exchange will change the proton hopping mechanism and the proton-hopping pathway.

6.5.3 Proton hopping mechanism

The changes of d1–d3 in different protonation states of the system (Figure 6a) as well as the dynamic evolution of d1–d3 in the two types of protonation equilibria (Figure 6b) indicate that the fluctuations of the hydrogen-bonding network involving the donor–acceptor pair are important driving forces for the activation of the proton transfer process. Another critical criterion for the hopping to occur is the distance between donor and acceptor, $R_{DA.}$ The hopping events happen only when R_{DA} and d1-d3 fulfill certain requirements. In vacuum, the protonated acetic acid is energetically more favorable than the hydronium ion at large $R_{DA.}$ At decreasing R_{DA}, the protonated acetic acid becomes less favorable. In aqueous solution environment, the well-established hydrogen-bonding network with neighboring water molecules

provides further stabilization of the hydronium ion. Fluctuations of this hydrogen-bonding network weaken this effect and facilitate the back transfer of the proton to acetic acid.

In the previous studies of hydrated excess proton in water using either the CPMD (*22*) method or the MS-EVB (*55*) model, the mechanism of proton hopping between closest water molecules was proposed as the fluctuation-induced breakage of a hydrogen bond between the first and second solvation shell of hydronium ion (*22*). The fluctuation of the hydrogen-bonding network in some stages destabilizes the donor (hydronium ion) and stabilizes the protonated acceptor, which results in the proton transfer events (*55*). The transfer statistics observed in our simulation agrees well with this mechanism. Moreover, the changes of R_{DA} in our simulation may also be coupled to the fluctuations of the hydrogen-bonding network. Another evidence for the above mentioned mechanism are the hopping events at low E_{12} when the proton hops from hydronium ion to acetic acid (see Figure 6c). It is clear that the low E_{12} is due to the weakening of the hydrogen-bonding network as well as due to the separation of the donor–acceptor pair.

6.6 Conclusion

For the first time, the dynamic protonation equilibrium between an amino acid side chain analogue and bulk water was successfully observed through unbiased computer simulations. Two different behaviors of the proton transfer were identified during a 50 ns Q-HOP MD simulation. During the fast-swapping equilibrium that occupied most of the simulation time, the proton mostly stayed on the acetic acid. Several times per nanosecond the proton left the acetic acid, was then exchanged between different water molecules, and finally came back to the acetic acid. The fluctuations of the hydrogen-bonding network as well as the donor acceptor distance were found to be the driving force of the activated processes. The pK_a of acetic acid calculated based on the relative population of protonated and deprotonated states and the diffusion coefficient of the excess proton agree well with the experimental measurements.

References

1. Warshel, A., Narayszabo, G., Sussman, F., and Hwang, J. K. (1989) How Do Serine Proteases Really Work, *Biochemistry 28*, 3629-3637.
2. Li, G. S., Maigret, B., Rinaldi, D., and Ruiz-Lopez, M. F. (1998) Influence of environment on proton-transfer mechanisms in model triads from theoretical calculations, *J. Comput. Chem. 19*, 1675-1688.
3. Umeyama, H., Hirono, S., and Nakagawa, S. (1984) Charge State of His-57-Asp-102 Couple in a Transition-State Analog Trypsin Complex - a Molecular-Orbital Study, *Proc. Natl. Acad. Sci. U. S. A. 81*, 6266-6270.
4. Lu, D. S., and Voth, G. A. (1998) Proton transfer in the enzyme carbonic anhydrase: An ab initio study, *J. Am. Chem. Soc. 120*, 4006-4014.
5. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P., and Lanyi, J. K. (1999) Structural changes in bacteriorhodopsin during ion transport at 2 Angstrom resolution, *Science 286*, 255-260.
6. Iwata, S., Ostermeier, C., Ludwig, B., and Michel, H. (1995) Structure at 2.8-Angstrom Resolution of Cytochrome-C-Oxidase from Paracoccus-Denitrificans, *Nature 376*, 660-669.
7. Michel, H. (1999) Cytochrome c oxidase: Catalytic cycle and mechanisms of proton pumping-A discussion, *Biochemistry 38*, 15129-15140.
8. Abrahams, J. P., Leslie, A. G. W., Lutter, R., and Walker, J. E. (1994) Structure at 2.8-Angstrom Resolution of F1-Atpase from Bovine Heart-Mitochondria, *Nature 370*, 621-628.
9. Dlugosz, M., and Antosiewicz, J. M. (2005) The impact of protonation equilibria on protein structure, *J. Phys.: Condens. Matter 17*, S1607-S1616.
10. Gu, W., Wang, T. T., Zhu, J., Shi, Y. Y., and Liu, H. Y. (2003) Molecular dynamics simulation of the unfolding of the human prion protein domain under low pH and high temperature conditions, *Biophys. Chem. 104*, 79-94.
11. Alonso, D. O. V., DeArmond, S. J., Cohen, F. E., and Daggett, V. (2001) Mapping the early steps in the pH-induced conformational conversion of the prion protein, *Proc. Natl. Acad. Sci. U. S. A. 98*, 2985-2989.
12. Lecoutre, J., Tittor, J., Oesterhelt, D., and Gerwert, K. (1995) Experimental-Evidence for Hydrogen-Bonded Network Proton-Transfer in Bacteriorhodopsin Shown by Fourier-Transform Infrared-Spectroscopy Using Azide as Catalyst, *Proc. Natl. Acad. Sci. U. S. A. 92*, 4962-4966.
13. Borjesson, U., and Hunenberger, P. H. (2001) Explicit-solvent molecular dynamics simulation at constant pH: Methodology and application to small amines, *J. Chem. Phys. 114*, 9706-9719.
14. Baptista, A. M., Martel, P. J., and Petersen, S. B. (1997) Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration, *Proteins 27*, 523-544.
15. Onufriev, A., Case, D. A., and Ullmann, G. M. (2001) A novel view of pH titration in biomolecules, *Biochemistry 40*, 3413-3419.
16. Mongan, J., Case, D. A., and McCammon, J. A. (2004) Constant pH molecular dynamics in generalized born implicit solvent, *J. Comput. Chem. 25*, 2038-2048.
17. Lee, M. S., Salsbury, F. R., and Brooks, C. L. (2004) Constant-pH molecular dynamics using continuous titration coordinates, *Proteins 56*, 738-752.
18. Walczak, A. M., and Antosiewicz, J. M. (2002) Langevin dynamics of proteins at constant pH, *Phys. Rev. E 66*.
19. Baptista, A. M., Teixeira, V. H., and Soares, C. M. (2002) Constant-pH molecular dynamics using stochastic titration, *J. Chem. Phys. 117*, 4184-4200.
20. Burgi, R., Kollman, P. A., and van Gunsteren, W. F. (2002) Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation, *Proteins 47*, 469-480.

21. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1997) On the quantum nature of the shared proton in hydrogen bonds, *Science 275*, 817-820.
22. Marx, D., Tuckerman, M. E., Hutter, J., and Parrinello, M. (1999) The nature of the hydrated excess proton in water, *Nature 397*, 601-604.
23. Marx, D., and Parrinello, M. (1996) Ab initio path integral molecular dynamics: Basic ideas, *J. Chem. Phys. 104*, 4077-4082.
24. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1996) Efficient and general algorithms for path integral Car-Parrinello molecular dynamics, *J. Chem. Phys. 104*, 5579-5588.
25. Lobaugh, J., and Voth, G. A. (1996) The quantum dynamics of an excess proton in water, *J. Chem. Phys. 104*, 2056-2069.
26. Cao, J. S., and Voth, G. A. (1994) The Formulation of Quantum-Statistical Mechanics Based on the Feynman Path Centroid Density .1. Equilibrium Properties, *J. Chem. Phys. 100*, 5093-5105.
27. Schmitt, U. W., and Voth, G. A. (1998) Multistate empirical valence bond model for proton transport in water, *J. Phys. Chem. B 102*, 5547-5551.
28. Schmitt, U. W., and Voth, G. A. (1999) The computer simulation of proton transport in water, *J. Chem. Phys. 111*, 9361-9381.
29. Day, T. J. F., Soudackov, A. V., Cuma, M., Schmitt, U. W., and Voth, G. A. (2002) A second generation multistate empirical valence bond model for proton transport in aqueous systems, *J. Chem. Phys. 117*, 5839-5849.
30. Maupin, C. M., Wong, K. F., Soudackov, A. V., Kim, S., and Voth, G. A. (2006) A multistate empirical valence bond description of protonatable amino acids, *J. Phys. Chem. A 110*, 631-639.
31. Pangali, C., Rao, M., and Berne, B. J. (1979) Monte-Carlo Simulation of the Hydrophobic Interaction, *J. Chem. Phys. 71*, 2975-2981.
32. Patey, G. N., and Valleau, J. P. (1975) Monte-Carlo Method for Obtaining Interionic Potential of Mean Force in Ionic Solution, *J. Chem. Phys. 63*, 2334-2339.
33. Izvekov, S., and Voth, G. A. (2005) Ab initio molecular-dynamics simulation of aqueous proton solvation and transport revisited, *J. Chem. Phys. 123*.
34. Lill, M. A., and Helms, V. (2002) Proton shuttle in green fluorescent protein studied by dynamic simulations, *Proc. Natl. Acad. Sci. U. S. A. 99*, 2778-2781.
35. Bondar, A. N., Elstner, M., Suhai, S., Smith, J. C., and Fischer, S. (2004) Mechanism of primary proton transfer in bacteriorhodopsin, *Structure 12*, 1281-1288.
36. Pomes, R., and Roux, B. (1996) Structure and dynamics of a proton wire: a theoretical study of H+ translocation along the single-file water chain in the gramicidin A channel, *Biophys. J. 71*, 19-39.
37. Pomes, R., and Roux, B. (1998) Free Energy Profiles for H+ Conduction along Hydrogen-Bonded Chains of Water Molecules, *Biophys. J. 75*, 33-40.
38. Xu, J. C., and Voth, G. A. (2005) Computer simulation of explicit proton translocation in cytochrome c oxidase: The D-pathway, *Proc. Natl. Acad. Sci. U. S. A. 102*, 6795-6800.
39. Braun-Sand, S., Strajbl, M., and Warshel, A. (2004) Studies of proton translocations in biological systems: Simulating proton transport in carbonic anhydrase by EVB-based models, *Biophys. J. 87*, 2221-2239.
40. Lill, M. A., and Helms, V. (2001) Reaction rates for proton transfer over small barriers and connection to transition state theory, *J. Chem. Phys. 115*, 7985-7992.
41. Lill, M. A., and Helms, V. (2001) Molecular dynamics simulation of proton transport with quantum mechanically derived proton hopping rates (Q-HOP MD), *J. Chem. Phys. 115*, 7993-8005.
42. Lill, M. A., and Helms, V. (2001) Compact parameter set for fast estimation of proton transfer rates, *J. Chem. Phys. 114*, 1125-1132.
43. Lill, M. A., Hutter, M. C., and Helms, V. (2000) Accounting for environmental effects in ab initio calculations of proton transfer barriers, *J. Phys. Chem. A 104*, 8283-8289.
44. de Groot, B. L., Frigato, T., Helms, V., and Grubmuller, H. (2003) The mechanism of proton exclusion in the aquaporin-1 water channel, *J. Mol. Biol.333*, 279-293.
45. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995), *J. Am. Chem. Soc. 118*, 2309-2309.

46. Herzog, E., Frigato, T., Helms, V., and Lancaster, C. R. D. (2006) Energy Barriers of Proton Transfer Reactions between Amino Acid Side Chain Analogs and Water from ab initio Calculations, *J. Comput. Chem. 27,* 1534-1547
47. Berendsen, H. J. C., Gigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials, *J. Phys. Chem. 91*, 6269-6271
48. Straatsma, T. P., Philippopoulos, M., and McCammon, J. A. (2000) NWChem: Exploiting parallelism in molecular simulations, *Comput. Phys. Commun. 128*, 377-385.
49. Kendall, R. A., Apra, E., Bernholdt, D. E., Bylaska, E. J., Dupuis, M., Fann, G. I., Harrison, R. J., Ju, J. L., Nichols, J. A., Nieplocha, J., Straatsma, T. P., Windus, T. L., and Wong, A. T. (2000) High performance computational chemistry: An overview of NWChem a distributed parallel application, *Comput. Phys. Commun. 128*, 260-283.
50. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A Smooth Particle Mesh Ewald Method, *J. Chem. Phys. 103*, 8577-8593.
51. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 684-3690.
52. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes, *J. Comput. Phys. 23*, 327-341.
53. Isaacson, A. D., and Truhlar, D. G. (1982) Polyatomic Canonical Variational Theory for Chemical-Reaction Rates - Separable-Mode Formalism with Application to Oh+H2- H2o+H, *J. Chem. Phys. 76*, 1380-1391.
54. Field, M. J., Bash, P. A., and Karplus, M. (1990) A Combined Quantum-Mechanical and Molecular Mechanical Potential for Molecular-Dynamics Simulations, *J. Comput. Chem. 11*, 700-733.
55. Lapid, H., Agmon, N., Petersen, M. K., and Voth, G. A. (2005) A bond-order analysis of the mechanism for hydrated proton mobility in liquid water, *J. Chem. Phys. 122*.

Chapter 7

Different Protonation Equilibria of 4-Methylimidazole and Acetic Acid

(accepted for publication in *ChemPhysChem* (2007))

7.1 Summary

Dynamic protonation equilibria in water of one 4-methylimidazole molecule as well as for pairs and groups consisting of 4-methylimidazole, acetic acid and bridging water molecules were studied using Q-HOP molecular dynamics simulation. We found a qualitatively different protonation behavior of 4-methylimidazole compared to that of acetic acid. On one hand, deprotonated, neutral 4-methylimidazole cannot as easily attract a freely diffusing extra proton from solution. Once the proton is bound, however, it remains tightly bound on a time scale of tens of nanoseconds. In a linear chain composed of acetic acid, a separating water molecule and 4-methylimidazole, an excess proton is equally shared between 4-methylimidazole and water. When a water molecule is linearly placed between two acetic acid molecules, the excess proton is always found on the central water. On the other hand, an excess proton in a 4-methylimidazole–water–4-methylimidazole chain is always localized on one of the two 4-methylimidazoles. These findings are of interest to the discussion of proton transfer along chains of amino acids and water molecules in biomolecules.

7.2 Introduction

Protonation equilibria are essential in many biological and chemical processes. For example, cellular proton pumps such as cytochrome *c* oxidase (*1, 2*) and cytochrome bc_1 complex (*3, 4*) generate proton gradients across biological membranes, which are then used by other biological processes such as for the synthesis of ATP. Proton transfer reactions are also crucial in other areas, for example, for membrane permeation in hydrogen fuel cells or in polymers (*5*). In spite of their enormous importance, many aspects of proton transfer (PT) reactions in biomolecules remain poorly understood. Due to the fundamental and / or technical difficulties with respect to the direct experimental observation of PT reactions, it is highly desirable to complement the existing experimental techniques by computational methods (*6, 7*).

In biological systems, especially in membrane proton pumps, the proton transfer pathways may extend over several nanometers involving many titratable amino acid as well as water chains (*8*). The mechanisms of such long transfer processes are, therefore, rather complicated (*2*) and the principles behind these long distance proton-transport processes are best revealed by transferring knowledge obtained on well-understood model systems or subprocesses of the more complex systems. In this regard, proton transfer in aqueous solution has been extensively studied in the past decade by ab initio molecular dynamics (AIMD) simulation (*9-12*), DFT-based molecular dynamics (*13*), empirical valence bond methods (*14-20*). based on the work by Warshel (*21*) and other approaches (*22-24*). Several studies also addressed proton transfer equilibria involving amino acids. For example, Car-Parrinello molecular dynamics (CPMD) (*25, 26*) combined with metadynamics and transition path sampling was employed to compute free energy profiles for the deprotonation of acetic acid in water (*27*). Klein and co-workers applied constrained AIMD (*28*) as well as CPMD combined with Potential of Mean Force (PMF) calculations (*29*) to study the deprotonation of histidine in water solution and calculated the relative pK_a value of deprotonation. A recent study presented the dynamic simulation of pK_a values for amino acids (*30*) using the MS-EVB model and the umbrella sampling technique (*31, 32*). The studies mentioned above provided quantum or semiempirical descriptions of short-range proton transfer process or described the diffusion of excess protons in bulk water. However, no such study has so far addressed the long term diffusion of protons involving both amino acids and solvent environment. Although several studies have successfully combined quantum mechanical and molecular mechanics force field approaches (QM/MM) with path-sampling techniques (*33-35*) to study proton transfer in biological macromolecules, there is certainly a great need for semi-quantitative approaches that can efficiently explore proton transfer paths in proteins consisting of many subsequent transfer events. Once these transfer paths are identified, more accurate and well-established methods can be applied to compute PMFs of individual reaction steps.

One such model of intermediate accuracy, the Q-HOP MD method, was introduced earlier to study proton transport in biomolecular systems (*36-39*). In the Q-HOP scheme, the dynamics of a classical simulation system is propagated by conventional newtonian molecular dynamics and stochastic proton transfer events allow for dynamic protonation changes of the titratable groups in the system. The corresponding proton transfer events are abstracted as quasi-instantaneous hopping from a donor group to an acceptor group. In this way, the total number of protons is conserved. The proton transfer likelihood per time step, termed the proton transfer probability, is calculated on the fly for each proton donor and acceptor pair using a parameterized functional form during the MD simulation. Depending on whether the proton transfer takes place or not (by comparing the proton transfer probability with a random number), the topology of the system will be modified or be kept unchanged before the next step of MD simulation. The transfer probabilities depend on the actual donor-acceptor distance, termed RDA, and the energy difference between the two minima at donor and acceptor, termed E_{12} in the momentary configuration. In contrast to MS-EVB and ab initio based molecular dynamics simulation methods, the Q-HOP method does not include an explicit treatment of a delocalized proton. These details are believed to be of less importance for identification of transfer pathways. In protonation equilibria as studied in this work, a shared proton between donor and acceptor would be reflected by frequent exchanges between both groups. The Q-HOP

MD method has been successfully applied to study the proton shuttle in green fluorescent protein (GFP) (*40*) and to understand the mechanism of proton blockage in Aquaporin (*41*). In a recent study, we presented its application to study the explicit protonation equilibrium of solvated acetic acid on a time scale of tens of nanoseconds at p*H* 1 (42). Two different protonation equilibria were observed through unbiased Q-HOP MD simulations and the pK_a of acetic acid was calculated based on the relative populations of protonated and deprotonated states observed during the simulation, see eq. (9) below. In the current study, similar calculations were performed for 4-methylimidazole (a side chain analog of histidine). In biological systems, titratable amino acids that may be involved in proton transfer pathways are either exposed to bulk solution, make direct contact with each other, or are separated by water molecule(s). To mimic these cases using model systems, Q-HOP simulations were also performed on pairs and groups consisting of 4-methylimidazole, acetic acid and water molecules in aqueous solution. Interestingly, we observed a qualitatively different protonation behavior of 4-methylimidazole compared to that of acetic acid. The results for all possible pairs/groups of 4-methylimidazole and acetic acid should be of interest to the discussion and understanding of proton transfer in large biomolecules.

7.3 Computational Methods

7.3.1 Q-HOP method

In the Q-HOP method, the proton hopping probability p is calculated from two quantities that can be readily computed on the fly from the momentary coordinates: the energy difference E_{12} between the proton transfer reaction products and reactants, and the distance between donor and acceptor atoms, R_{DA} (*38*), see equations (5)-(7) below. Depending on the values of E_{12} and R_{DA}, three different approaches based on precomputed parameterizations are used to compute the hopping probability (*36*). For large E_{12} and large R_{DA}, a modified expression from transition state theory is used that also accounts for the zero-point energy and tunneling effects:

$$p_{TST} = \kappa(T, E_M)\frac{k_B T}{h}\exp\left(-\frac{E_b - h\omega/2}{k_B T}\right)\Delta t \qquad (1)$$

Here, $\kappa(T, E_M)$ is the enhancement of the classical transfer rate due to tunneling $\kappa = k_{QM}/k_{classical}$ as a function of temperature T and E_M. $E_M = E_{\max} - \max(E_{\min,1}, E_{\min,2})$ is the difference between the energy maximum E_{max} and the higher one of the two energy minima $E_{min,1}$ and $E_{min,2}$ along the double well potential of a typical transition reaction. $h\omega/2$ is the zero-point energy obtained by considering the bonds that contain a transferring proton as a quantum-mechanical harmonic oscillator with frequency at the educt well minimum (*38*). E_b is the energy barrier along the one-dimensional double well potential and is calculated in Q-HOP as a function of E_{12} and R_{DA}:

$$E_b(E_{12}, R_{DA}) = S(R_{DA}) + T(R_{DA})E_{12} + V(R_{DA})E_{12}^2 \qquad (2)$$

where *S*, *T* and *V* have a simple functional dependence on R_{DA} (*38*). We note that eq. (1) should, strictly speaking, contain the activation free energy, not the energy of the barrier.

On the other hand, this high-barrier regime is only a limiting case. In nanosecond time-scale simulations, the probabilities computed from eq. (1) are too low for PT events to occur. However, situations arise during simulations, in which the reaction energy barrier is very small, or even vanishes due to occasional small donor-acceptor distances or due to the environmental influence on the reaction energy profile. In such cases of small energy barriers TST assumptions are known to break down. Following the discussion of Hänggi *et al.* (*43*), TST is valid in a strict sense if $\exp(E_b/k_BT) >> 1$. As explained in a previous work (*36*), we consider eq. (1) to be valid within the Q-HOP framework if $\exp(E_b(E_{12}, R_{DA})/k_BT) \geq 100$.

In situations where the TST validity criterion is not satisfied, i.e. for barriers involving small E_{12} and small R_{DA}, a different approach is used, and the hopping probability is estimated as follows:

$$p_{TDSE} = 0.5 \tanh(-K(R_{DA})E_{12} + M(R_{DA})) + 0.5 \tag{3}$$

with *K* and *M* being functions of R_{DA} (*36*). The fitting formula (3) was previously obtained (*36*) following the propagation of a one-dimensional wave packet as the solution of the time-dependent Schrödinger equation (TDSE) for many different potential curves characterized by different E_{12} and R_{DA} values, and computing the fractional population that crosses the barrier. Again, due to numerical noise in the wave packet propagation, a validity criterion was introduced, and eq. (3) is considered valid only if $p_{TDSE} > 0.1$ (*36*).

As the high energy and the low energy regimes do not overlap, a third intermediate regime was introduced (*36*): for given E_{12} and R_{DA}, for which neither TST nor TDSE validity criteria apply, hopping probabilities are obtained by interpolating linearly, on a logarithmic scale, the values on the borders of the two regimes previously defined:

$$\log_{10} p_{gap}(E_{12}) = \log_{10} p_{TDSE}(E_{12}^L) + \frac{\log_{10} p_{TST}(E_{12}^R) - \log_{10} p_{TDSE}(E_{12}^L)}{E_{12}^R - E_{12}^L}\left(E_{12} - E_{12}^L\right) \tag{4}$$

where E_{l2}^L and E_{l2}^R are the validity limits of the TDSE and TST approach respectively (*36*).

In the Q-HOP method, E_{12} is a sum of two contributions:

$$E_{12} = E_{12}^0 + E_{12}^{env} \tag{5}$$

E_{12}^0 is the energy difference between the reaction products and reactants in vacuum. It is obtained from the following empirical relationship (*38*),

$$E_{12}^0 = \alpha + \beta \cdot R_{DA} + \gamma \cdot R_{DA}^2 \tag{6}$$

where *a*, *b* and *g* are fitted parameters compiled in a recent data set based on MP2/6-31++G** calculations of all titratable amino acids (*44*). The values for the donor-acceptor pairs considered in this study were taken from this previous work (*44*). The environmental contribution E_{12}^{env} is calculated from the coulombic interactions between the two pairs and the environment:

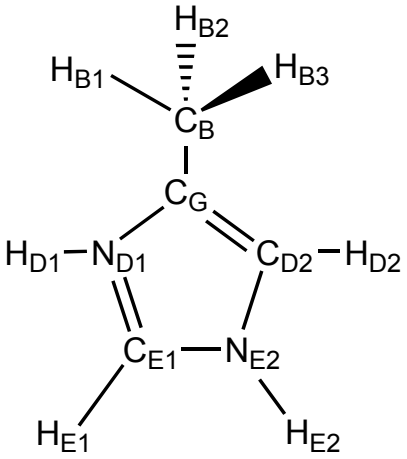
$$E_{12}^{env} = E_{DH-A,env}^{Coul} + E_{D-AH,env}^{Coul} \tag{7}$$

where $E_{DH-A,env}^{Coul}$ and $E_{D-AH,env}^{Coul}$ are defined as:

$$E^{Coul} = \sum_{i}^{donor\_acceptor\_atoms} \sum_{j}^{remaining\_system} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}}. \tag{8}$$

q_i and q_j are the respective atomic partial charges, and r_{ij} is the atomic distance between atoms *i* and *j*. The index *i* runs over the donor-acceptor pair atoms, and the index *j* over the remaining atoms. ε_0 is the permittivity of vacuum. First, E_{12} was calculated by equation (5)-(7) for 100 equally spaced snapshots taken from a MD simulation of protonated and deprotonated 4-methylimidazole in a solvent box of SPC/E water molecules (*45*) using the atomic partial charges from the AMBER force field. As in our previous work (*42*) on solvated acetic acid, the computed E_{12} values showed systematic deviations from reference QM/MM calculations of between 8 and 25 kcal/mol (data not shown). Therefore, we optimized a separate set of q_{12}^{env} charges to reproduce the results from the QM/MM calculations (*42*). A similar refinement was recently introduced in the MS-EVB model as well (*30*). We stress that these charges are not used in the propagation of the trajectory (normal MD part), but only serve to compute the environmental electrostatic contribution from equation (7).

7.3.2 Simulation setup



Scheme 1: Topology of 4-methylimidazole

Atom	Partial atomic charges (e) during standard MD			Partial atomic charges q_{12}^{env} (e) to compute E_{12}^{env}		
	State1[a]	State2	State3	State1	State2	State3
HB (1-3)	0.0810	0.0367	0.0402	0.0810	0.0367	0.0402
CB	-0.0651	-0.0806	-0.0939	-0.0651	-0.0806	-0.0939
CG	-0.0012	0.1868	-0.0266	-0.0012	0.1868	-0.0266
ND1	-0.1513	-0.5432	-0.3811	-0.6513	-0.5432	-0.3811
HD1	0.3866	dummy	0.3649	0.8866	dummy	0.3649
CD2	-0.1141	-0.2207	0.1292	-0.1141	-0.2207	0.1292
HD2	0.2317	0.1862	0.1147	0.2317	0.1862	0.1147
CE1	-0.0170	0.1635	0.2057	-0.0170	0.1635	0.2057
HE1	0.2681	0.1435	0.1392	0.2681	0.1435	0.1392
NE2	-0.1718	-0.2795	-0.5727	-0.6718	-0.2795	-0.5727
HE2	0.3911	0.3339	dummy	0.8911	0.3339	dummy

Table 1 Partial atomic charges for 4-methylimidazole. [a] State1: protonated on both ND1 and NE2; State2: protonated on NE2; State3: protonated on ND1.

The segments of protonated and deprotonated 4-methylimidazole (in this paper referred to as $4MIH^+$ and 4MI) were constructed based on the segments of protonated and deprotonated histidine in the AMBER force field (*46*). The C_α–C_β bond was deleted and the C_α atom was replaced by an H_β atom. All bonded and non-bonded parameters of the newly added H_β atom were set to the same values as for the other two H_β atoms. The remaining atomic partial charge was added to the C_β atom to make the net charge of the segment integer (0 for 4MI and +1 for $4MIH^+$). The parameterization of 4-methylimidazole used in this study is listed in detail in Scheme 1 and Table 1. The segments of protonated and deprotonated acetic acid (alias as ACH and AC^-) were taken from our previous study (*42*).

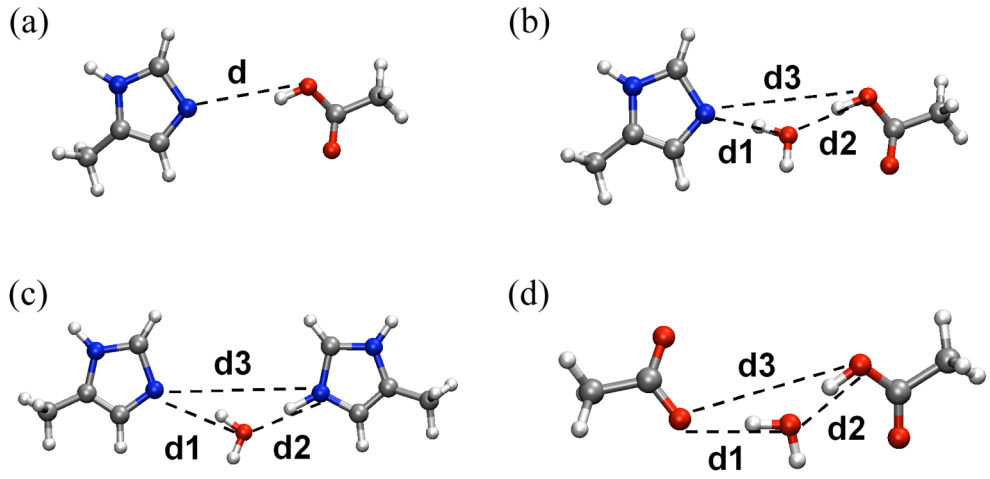


Figure 1: Schematic representations of the distance restraints employed during the simulations of (a) the 4MI–ACH pair, *d* was restrained to within 2.2 to 2.8 Å; (b) the 4MI–H_2O–ACH group, *d1* and *d2* were restrained to within 2.4 to 2.8 Å and *d3* was restrained at 4.6±0.2, 4.8±0.2, 5.0±0.2, 5.2±0.2 and 5.4±0.2 Å; (c) the 4MI–H_2O–$4MIH^+$ group, *d1* and *d2* were restrained to within 2.4 to 2.8 Å and *d3* was restrained at 4.6±0.2, 4.8±0.2, 5.0±0.2, 5.2±0.2 and 5.4±0.2 Å; (d) the AC^-–H_2O–ACH group, *d1* and *d2* were restrained to within 2.4 to 2.8 Å and *d3* was restrained at 4.6±0.2, 4.8±0.2, 5.0±0.2, 5.2±0.2 and 5.4±0.2 Å.

The Q-HOP MD simulations as well as the single point calculations for charge fitting were performed using a modified version of the NWChem 4.7 package employing the AMBER99 force field (*46*). In this implementation, the Particle Mesh Ewald (PME) method (*47*) is used for calculating long-range electrostatic interactions during molecular dynamics. The charge fitting of 4-methylimidazole followed the same procedure as in our previous study on acetic acid (*42*). For evaluating E_{12}^{env}, all coulombic interactions were computed between the donor–acceptor pairs and the other atoms of the simulation box. In the Q-HOP MD simulations, the $4MIH^+$–H_2O pair and the 4MI–H_3O^+ pair each were solvated in cubic boxes of 24 Å side length, using SPC/E water molecules (*45*). The 4MI–ACH pair, the 4MI–H_2O–ACH group, the 4MI–H_2O–$4MIH^+$ group and the AC^-–H_2O–ACH group were solvated in cubic boxes of 22 Å side length of SPC/E water molecules. All coordinate sets were first subjected to 500 steps of steepest-descent energy minimization (Q-HOP switched off). The solvent and modeled residues were then relaxed during a 100 ps MD simulation at 300 K prior to the Q-HOP MD simulation. For the $4MIH^+$–H_2O pair, the 4MI–H_3O^+ pair and the 4MI–ACH pair 10 ns Q-HOP MD simulations were performed on each system. In order to achieve adequate sampling of the protonation equilibrium for the 4MI–ACH pair on a feasible simulation time scale (ns to tens of ns), a distance restraint was applied between the atom $O_{\delta 1}$ of ACH and the atom $N_{\varepsilon 2}$ of 4MI (see Figure 1) using a harmonic potential with a force constant of 5.0×10^4 kJ/mol·nm^2. For the remaining systems, 2 ns Q-HOP MD simulations were carried out using different distance restraint schemes (see Figure 1 for details). In total, 60 ns of MD simulations were performed. During all simulations, temperature (300 K) and pressure (1 atm) were maintained by weak coupling to an external bath (*48*). The SHAKE procedure (*49*) was applied to constrain all bonds that contain hydrogen

atoms. Non-bonded interactions were treated using a cutoff of 9 Å and long-range electrostatic interactions were computed using the PME method. The time step of the simulations was 1 fs throughout. Scanning for possible proton transfer events was performed every 10 steps and snapshots were also recorded every 10 steps to track all hopping events. For the $4MIH^+$–H_2O pair and the 4MI–H_3O^+ pair, all water molecules as well as the 4-methylimidazole were possible donor/acceptors. All protons of the water molecules were transferable. In the simulations of the other pairs and groups, only the acetic acid, 4-methylimidazole and the bridging water molecule were possible donor/acceptors.

7.4 Results and Discussion

7.4.1 Solvated 4-methylimidazole

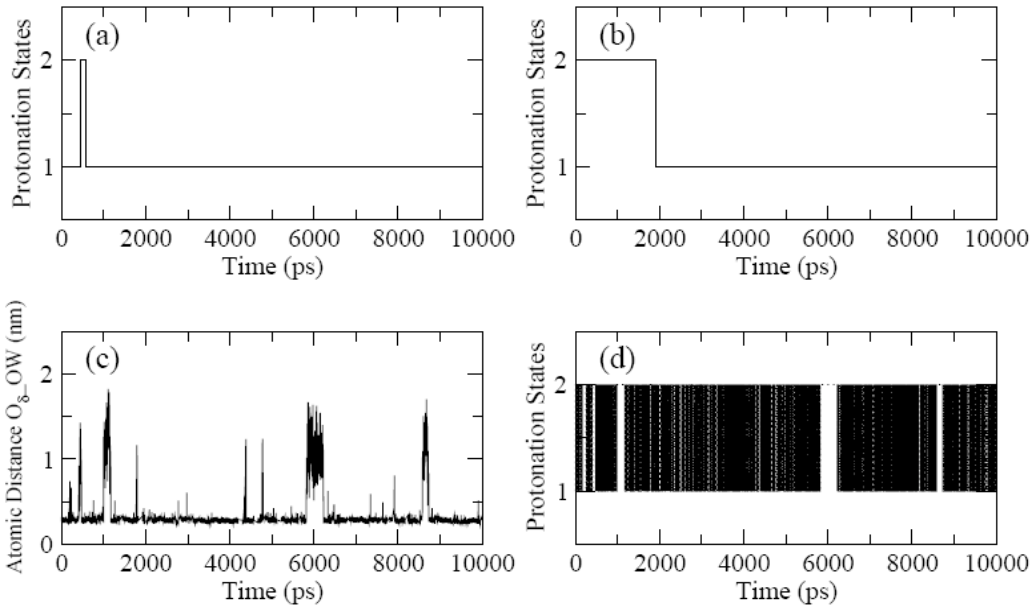


Figure 2: Time evolution of the protonation states in (a) the simulation of the $4MIH^+$–H_2O pair, where the acceptor atoms are $N_{\delta 1}$ or $N_{\varepsilon 2}$; and (b) in the simulation of the 4MI–H_3O^+ pair, where the acceptor atoms are $N_{\delta 1}$ or $N_{\varepsilon 2}$. 1 denotes the state in which the proton is located on the analog, 2 denotes the state in which the proton is bound to the hydronium ion. For comparison, (c) and (d) show corresponding data from acetic acid taken from ref 42. (c) Time evolution of the atomic distance between the acceptor atoms acetic acid and the OW atom of the hydronium ion; (d) time evolution of the protonation states of the systems in simulation of the solvated acetic acid.

In the simulations of solvated 4-methylimidazole (both $4MIH^+$–H_2O pair and the 4MI–H_3O^+ pair), extremely few transitions between $4MIH^+$(4MI) and H_2O (H_3O^+) were observed (see Figures 2a and 2b). This is in clear contrast to our previous study

on solvated acetic acid (*42*) where two different protonation equilibria were observed: a "fast-swapping" equilibrium where the protonated acetic acid shares the proton with a hydrogen-bonded water molecule; and a "travelling phase" when the proton escapes into the bulk volume and needs to rebind in the capture region of ca. 5 Å radius around acetic acid (see Figure 2c). As another noticeable difference compared to acetic acid, only one "travelling phase" of the H_3O^+ was observed in each simulation of 4-methylimidazole, whereas such "travelling phases" happened several times per nanosecond in the solvated acetic acid system. In contrast, the travelling time of H_3O^+ was longer than that found in the study of solvated acetic acid. The reason for the absence of the "fast swapping" equilibrium is the larger energy difference E_{12} between $4MIH^+$ and H_3O^+ in aqueous solution (see Figure 3a). That is about 10 kcal/mol larger on average than that between ACH and H_3O^+. Similar results were reported in a recent study using the MS-EVB method (*7*), where a difference of about 6–7 kcal/mol between the PMF for the proton abstraction from histidine and glutamic acid was found. This can also explain the lower frequency of "travelling phases" of H_3O^+, as it is very unlikely for $4MIH^+$ to be deprotonated at this low p*H*. On the other hand, as deprotonated 4MI has a neutral charge, a free H_3O^+ is not attracted by 4MI unless they come very close to each other. In contrast, deprotonated AC^- is negatively charged, and therefore attracts the positively charged H_3O^+ from longer distances. In other words, deprotonated AC^- has a larger capture radius for H_3O^+ than deprotonated 4MI. This is reflected by the radial distribution of H_3O^+ around the receptor atoms of 4MI and AC^- (see Figure 3b), where a peak at 4–5 Å in the solvated ACH/AC^- system (see Figure 2b) indicates the H_3O^+ capture radius of AC^- while no clear peak was found in the solvated $4MIH^+$/4MI system.

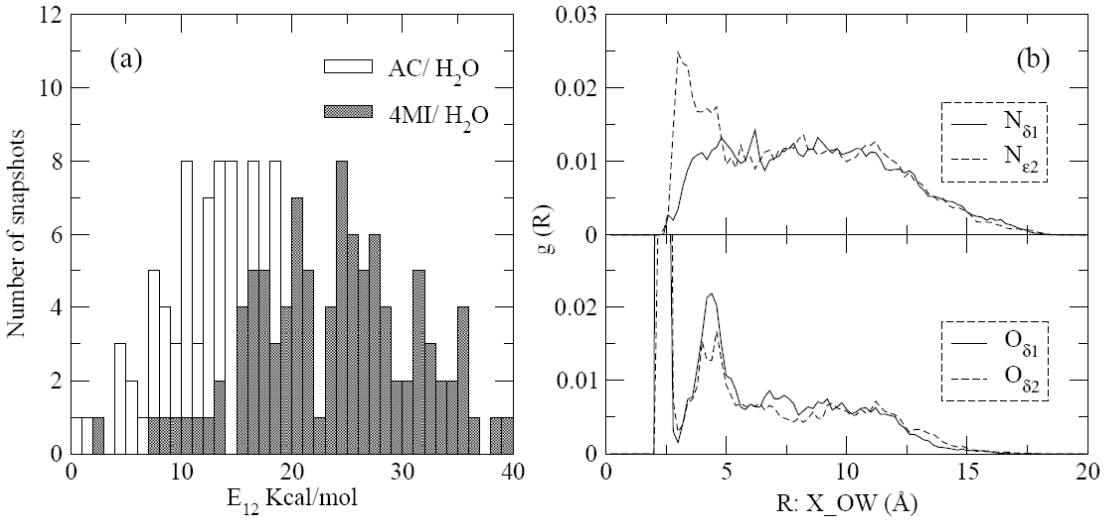


Figure 3: (a) The distribution of E_{12} between 4MI and H_2O and between AC and H_2O computed from Q-HOP for the 100 snapshots from a MD simulation; (b) Normalized radial distribution of hydronium ions around the two nitrogen atoms of 4MI (upper panel) and around the two carboxyl oxygen atoms of acetic acid (lower panel, data were taken from ref 42 for comparison).

The previously mentioned MS-EVB study of histidine and glutamic acid in aqueous solution (*7*) reported the PMF for proton dissociation from histidine and glutamic acid. Besides the metastable contact ion pair, the PMF curve of the glutamic

acid system for the location of the center of excess charge is funnel-shaped between 3 Å and 5 Å. The depth of this funnel is about 1–3 kcal/mol. This is an energetic analogue of the large proton capture radius of acetic acid that we found in the radial distribution of H_3O^+ (*42*). For the histidine system, however, the PMF curve obtained from MS-EVB simulations is rather flat indicating no obvious proton capture effect of histidine similar to what we found in the Q-HOP simulations reported here.

In our previous study of acetic acid, we calculated the pK_a value of acetic acid based on the relative populations of protonated and deprotonated acetic acid using the classical definition of pK_a:

$$pK_a = -\log\frac{[H^+][A^-]}{[HA]} \tag{9}$$

The calculated value of 3.0 was in quite good agreement with the experimental value of 4.7 (*42*). For that system, adequate sampling of the protonation equilibrium allowed us to directly convert the observed frequencies into a pK_a value. However, in the simulation of solvated $4MIH^+$/4MI, only a very limited number of exchange events were observed. We estimate that microseconds of simulation time would be required to achieve sufficient sampling for this system, which is currently not feasible. Therefore, calculating the pK_a value of $4MIH^+$ directly from eq. (9) is not valid at this point.

7.4.2 4MI–ACH pair

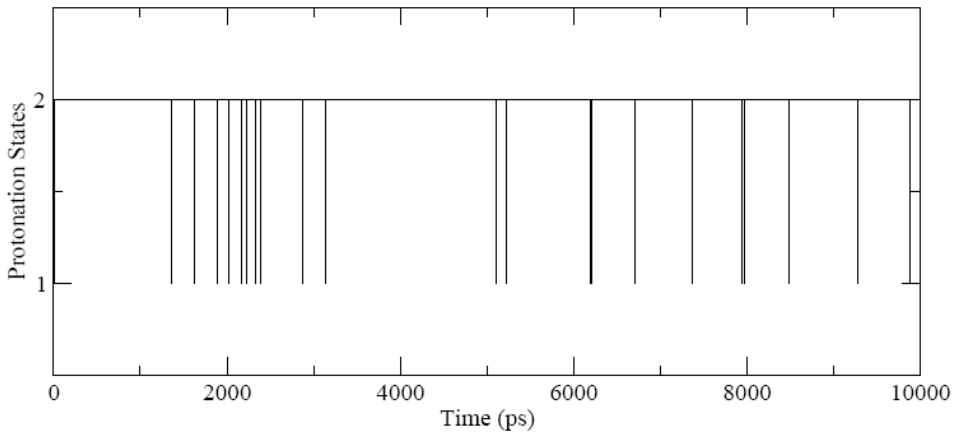


Figure 4: Time evolution of the protonation states of the system during the simulation of the 4MI–ACH pair. 1 denotes the state in which the proton is located on ACH, 2 denotes the state in which the proton is bound to $4MIH^+$.

To address the relative preference for $4MIH^+$ and ACH, a simulation of a solvated 4MI–ACH pair was performed employing a distance constraint between one of the carboxyl oxygen atoms and the N_ε atom of 4-methylimidazole. Figure 4 shows the protonation states of this pair during the 10 ns simulation. In total, 56 proton transfer

events occurred, 28 in each direction (proton exchange events). Protonated $4MIH^+$ occupied 99.9% of the full simulation time, whereas protonated ACH only occupied 0.1%. These occupancies indicate that $4MIH^+$ is about three orders of magnitude more favorable than ACH in aqueous solution. For comparison, the experimental pK_a measurement is 7.35 for 4-methylimidazole (*50*). However, the computed population ratio cannot be readily converted into a pK_a, since this simulation set up only reflects the pK_a of 4-methylimidazole at close distance from acetic acid with an artificial distance restraint applied. In an ideal scenario, 4MI and ACH should be fully solvated and at large distance from each other.

7.4.3 4MI–H_2O–ACH group

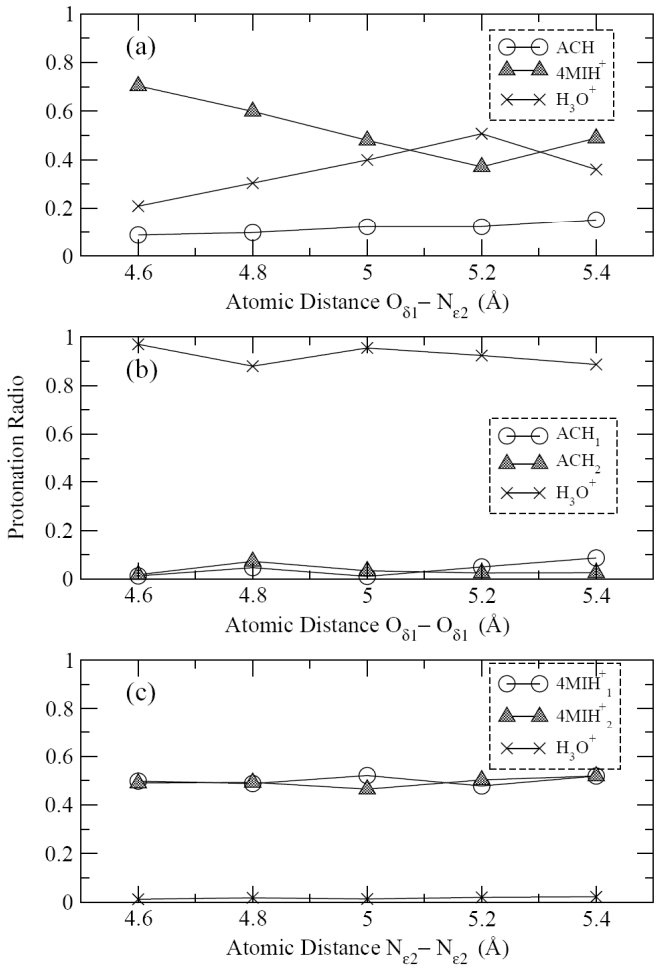


Figure 5: Protonation ratio during the simulations of (a) the 4MI–H_2O–ACH group; (b) the AC^-–H_2O–ACH group and (d) the 4MI–H_2O–$4MIH^+$ group.

As a next step, a bridging water molecule was placed in between the two groups. In order to achieve sufficient sampling of the proton equilibrium in this system, distance restraints were applied again (see Figure 1 for detail). While this is certainly artificial in some sense, such cases are often found in linear proton transfer chains in biological

macromolecules. In the solvated $4MIH^+$/4MI system as well as in the solvated 4MI–ACH pair, protonated $4MIH^+$ was predominantly found during the simulations. However, when placing a bridging water molecule between the 4MI-ACH pair, $4MIH^+$ shared the proton with the bridging water molecule (see Figure 5a). As mentioned before, in the Q-HOP scheme, proton sharing is reflected by frequent transfers between both groups and a similar residence time on both sides. This behavior of $4MIH^+$ only weakly depended on the distance restraint setup. $4MIH^+$ and the bridging water molecule shared the excess proton for 90% of the simulation, whereas ACH only occupied about 10% of the simulation time. At closer 4MI–ACH distance, $4MIH^+$ held the proton more often than the bridging water. For an enlarged distance, the occupancies of the proton converged to a half by half situation. During the simulations, 200 to 750 proton hopping events were observed depending on the restraining distance, indicating an adequate sampling of the protonation equilibria. The non-monotonous behavior between 5.0 Å and 5.4 Å is likely within the statistical error of the calculations. The detailed time evolution of the protonation states are shown in Figure 6.

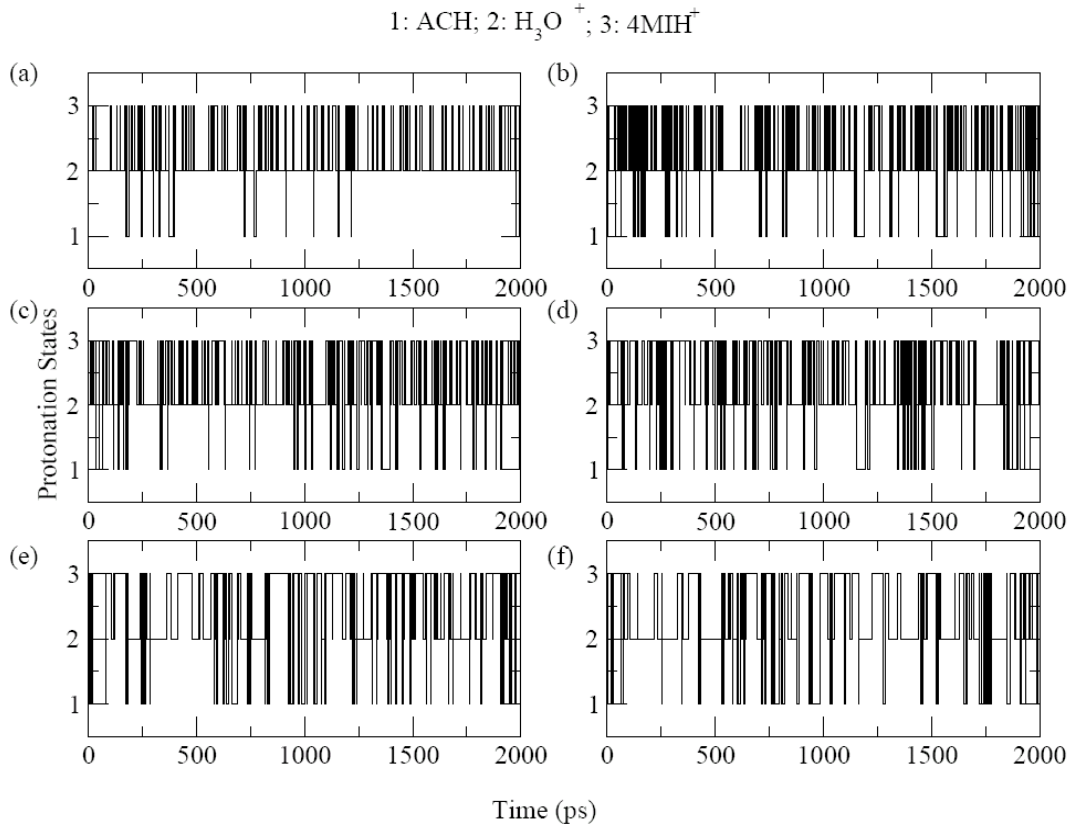


Figure 6: Detailed time evolution of the protonation states during the simulations of the 4MI-H_2O-ACH group. (a) No distance restraint was applied for d3 (see Figure 1 of the manuscript for descriptions). From (b) to (f), d3 was restrained at 4.6±0.2, 4.8±0.2, 5.0±0.2, 5.2±0.2 and 5.4±0.2 Å.

7.4.4 AC^-–H_2O–ACH group

When two acetic acid molecules were separated by a bridging water molecule, proton sharing was dominant. As shown in Figure 5b, the bridging water in fact carried the proton for most of the simulation time (about 90%), although many proton exchanges took place (about 130 to 200 hopping events in each simulation). This protonation behavior indicates that AC^- can transfer protons rapidly in a proton transport chain.

7.4.5 4MI–H_2O–$4MIH^+$ group

To complete the picture, two 4MI/$4MIH^+$ were simulated with a bridging water molecule in between. Interestingly, although proton sharing took place as in the other simulations, the protonation states of this group were completely different. During most of the simulation time (>97%), the proton stayed on either one of the two 4MI/$4MIH^+$. The water molecule held the proton for only very short times (see Figure 5c). In total, 150 to 350 hopping events were observed in each simulation. This sharply different behaviour of $4MIH^+$ illustrates that 4MI can serve as a good proton "container" to keep the proton. If the environment is favourable, however, $4MIH^+$ can also transfer the proton to its next destination. Cui and co-workers recently used DFT-TB molecular dynamics to characterize the free energy for proton transfer between a pair of 4MI separated a chain of three linear water molecules (*13*). They found that for groups with pK_a equal or larger than 7, a hole transfer mechanism becomes favorable. In particular, for transfer between two 4MI, transfer involving a hydroxide was about 1.2 kcal/mol for favorable than transfer involving hydronium (Grotthuss mechanism). Here, we did not account for the possibility of forming a hydroxide ion since we do not have a working Q-HOP parametrization for OH- in bulk water that reproduces the bulk proton diffusion constant. It would be interesting to apply the methodology of Riccardi et al. to the systems investigated here.

7.4.6 Biological insights

The Q-HOP MD simulations revealed that: 1) $HISH^+$ is very likely to keep the proton because of its relatively high pK_a. A close titratable group with lower pK_a only has few chances to snatch the proton from $HISH^+$ that are driven by environmental fluctuations. On the other hand, HIS has a relatively small proton capture radius, which makes it very hard to attract protons from long distances. 2) ASPH can easily share the proton with close titratable groups even if the acceptor group has a low pK_a. Moreover, ASP^- has a large proton capture radius (about 5 Å), which makes it an ideal proton "capturer". Can we relate these properties of ASP and HIS to one of their biological functions in terms of proton transport? We therefore analyzed the crystal structures of several typical proton pumps (*51-53*), Na+/H+ antiporter (*54*) as well as some non-proton related channels (ion channel, aquaporin) (*55, 56*).

7.4.6.1 Cytochrome C oxidase (COX) from Paracoccus denitrificans

COX is located in the inner membrane of mitochondria and many bacteria (*51, 57*) and catalyzes the terminal step of cellular respiration. During this step, four electrons are transferred from cytochrome *c* to dioxygen and four protons are transported across the located membrane. The X-ray structure of COX from *Paracoccus denitrificans*

(PDB entry: 1AR1) (*1*) is shown in Figure 7a. The zoomed view of the entrance of "D-pathway" (a proposed proton transfer pathway, proved by site-directed mutagenesis experiments (*51*)) shows the Asp (I-124)–HIS (I-28) pair. Continuum electrostatic calculations on the D-pathway of COX showed that Asp (I-124), His (I-28) and Asp (I-30) formed a strongly coupled cluster at the entrance of D-pathway (*8*).

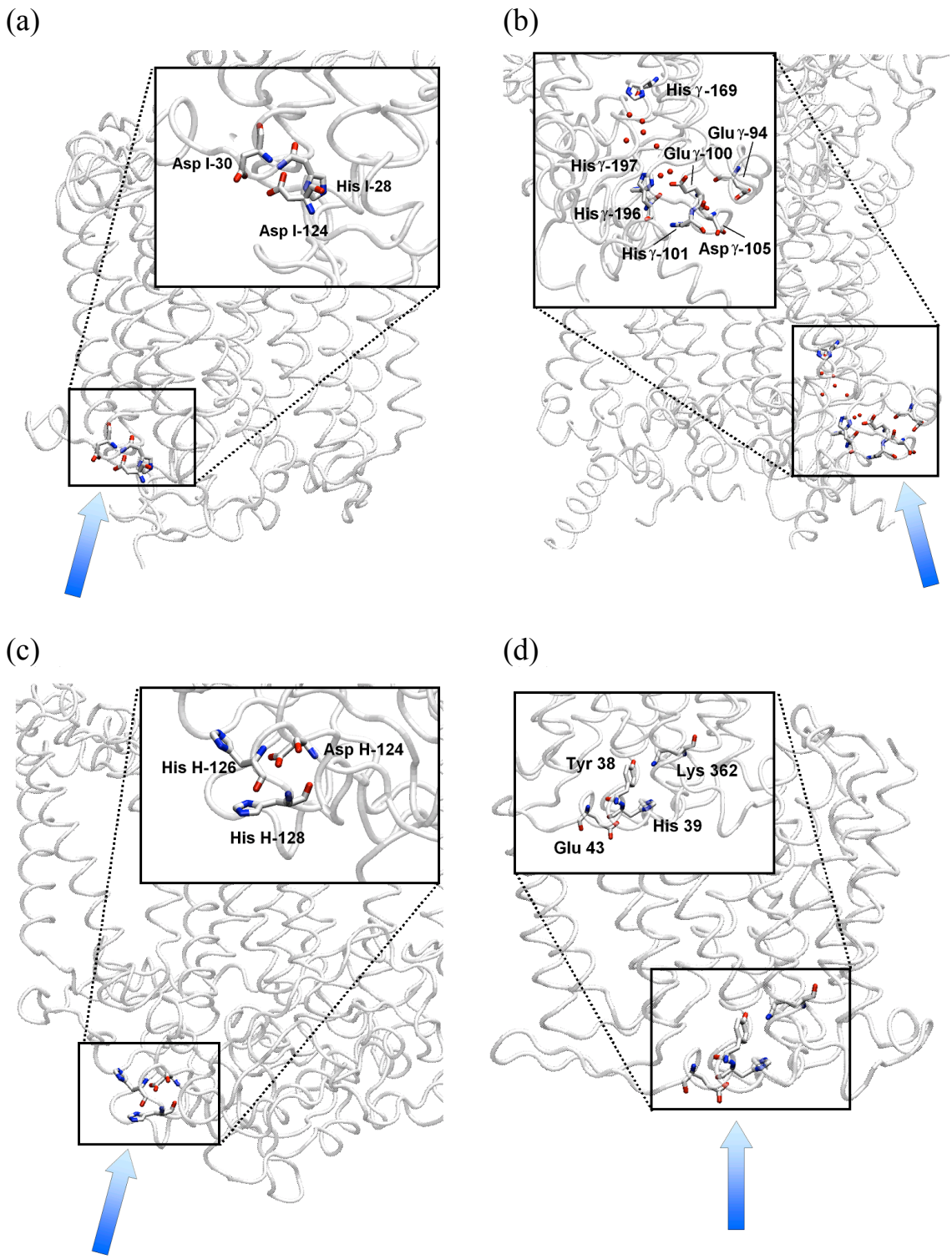


Figure 7: X-ray crystallographic structures of (a) Cytochrome *C* oxidase from *Paracoccus denitrificans*; (b) Formate dehydrogenase-N; (c) Bacterial reaction center and (d) Na^+ /H^+ antiporter from *Escherichia coli*. The peptide backbones of each protein are represented as tubes. The residues belonging to the clusters at putative proton entry sites discussed in the text are represented as sticks. Zoomed views of the proton entrances are shown in the insets. Important crystal water molecules are represented as small red balls. The blue arrows indicate the possible proton entrances suggested in the literature (see text).

7.4.6.2 Formate dehydrogenase-N (Fdh-N)

Fdh-N is an important part of the formate-nitrate oxidoreductase, which catalyzes the formate oxidation coupled nitrate respiration in *Escherichia coli* when oxygen is deficient (*58*). During the respiration reaction, two protons are translocated from the cytoplasmic side to the periplasmic side. X-ray crystallographic study on Fdh-N from *Escherichia coli* (see Figure 7b) suggested that a proton should be taken up from His (γ-169) to O_1 of the menaquinone (HQNO) when the second electron is transferred (*52*). However His (γ-169) is buried inside the protein. Therefore the deprotonated His (γ-169) needs to take up a proton from the cytoplasmic side via an extended water channel (*52*). Interestingly, the entrance of this water channel is lined by a cluster of Asp, Glu and His residues (see Figure 7b, PDB entry: 1QKF).

7.4.6.3 Bacterial reaction center (RC)

RC initiates the conversion of light into chemical energy in photosynthetic bacteria (*53*). RC from *Rhodobacter sphaeroides* (*59*) carries out this process through the reduction and protonation of a bound quinone molecule Q_B (the secondary quinone electron acceptor). In the RC embedded in the bacterial membrane, protons are taken up from the cytoplasm to form quinol at the Q_B site (*60*). Experimental studies on the inhibition of proton transfer showed that the binding of Zn^{2+} or Cd^{2+} to the surface of RC, (His (H-126), His (H-128) and Asp (H-124)) results in a dramatic reduction (more than 100-fold) in the proton transfer rate (*61*). *Paddock et al.* suggested this region as the site of proton entry (*61*) (see Figure 7c, PDB entry: 1AIJ (*62*)).

7.4.6.4 Na^+ /H^+ antiporter from Escherichia coli (NahA)

NahA is the main Na^+ /H^+ antiporter of *Escherichia coli* and many enterobacteria (*63*). It excretes Na^+ to the periplasm side in exchange for a backflow of protons into the cell using the proton gradient maintained across the bacterial membrane (*63*). Through this process, intracellular p*H*, cellular Na^+ content and cell volume are regulated for the living of cells. Multiconformation continuum electrostatic analysis of NahA showed that at the periplasmic funnel, where protons possibly enter, a cluster is formed by Tyr (38), His (39), Glu (43) and Lys (362) (*64*) (see Figure 7d, PDB entry: 1ZCD (*54*)).

7.4.6.5 Non-proton related channels

In ion channels (e.g. Na^+ / K^+ channel) (*55*) and aquaporin (*56*) that transmit ions or water molecules instead of protons, the entry points of respective substrate are mostly located at the centers of the proteins. Residues like His (Lys, Arg) Asp, Glu can also be found close to the entrances. However, we did not find the above-mentioned pair or cluster at the substrate entry sites of the proteins.

By inspecting crystal structures of several transmembrane proteins involved or not involved in proton translocation, we learnt that two types of amino acid residues (type 1: Asp and Glu; type 2: His, Lys and Arg) always appear together at the entrance of protons. Does this co-appearance tell us something about the very first step of all these long-range (e.g. across the membrane) proton transfer processes in membrane proteins? We will next consider the step even before entering the protein — when the protons are still diffusing outside the protein.

7.4.6.6 Proton diffusion in the presents of membranes

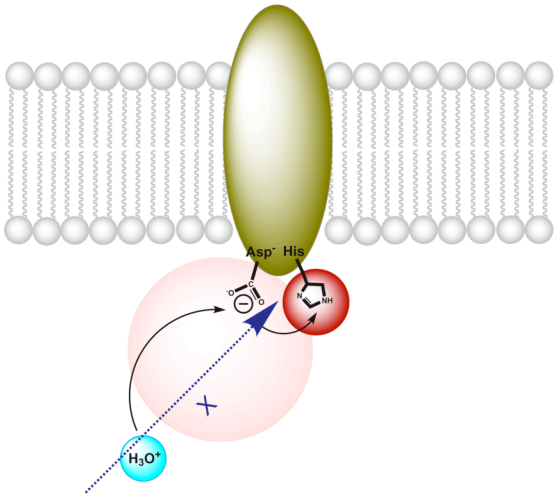
A free proton (or hydronium ion in aqueous solution) diffuses in bulk solution in a three dimensional manner. For situations close to a membrane, two different models were suggested in the past two decades: a "delocalized" model claims that protons are equilibrated between the proton transporter and the bulk phase before being taken up (*65*); a "localized" model asserts that protons move exclusively along the membrane surface and this diffusion is much faster since the protons now diffuse only in two dimensions (*66, 67*). In spite of these different models, a possible function of the co-appearance of type 1 and type 2 residues can be attributed under either framework.

According to the "delocalized" model, protons are partitioned between the protein and bulk solution. Considering the protonation properties of Asp (type 1) and His (type 2), type 1 residues are needed to attract protons from long distance with their large proton capture radii. To prevent the proton from being released back to the bulk, type 2 residues are needed to lock the proton on the surface of the protein or even inside the protein (see Figure 8a). Their roles in proton uptaking from the environment are not exchangeable because of the low pK_a of type 1 residues and the small proton capture radii of type 2 residues. In this model, these two types of residues perform concerted actions, namely "capture and store", at the site of proton entry of transmembrane proteins.

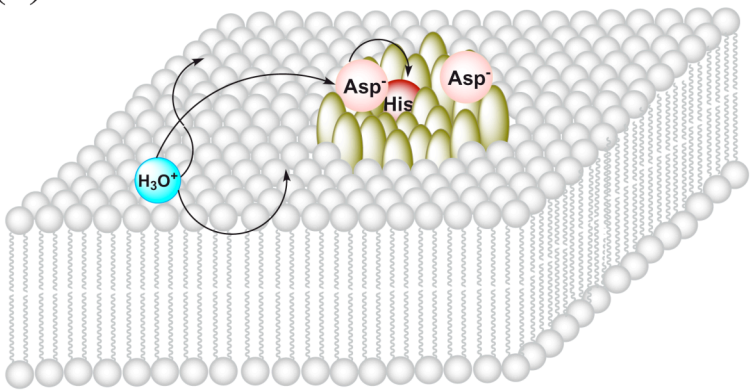
According to the "localized" model, protons diffuse along the surface of membrane. The details of the proton-conducing pathway along the membrane surface are still unclear. Molecular dynamics simulations employing the MS-EVB model of proton transport near the surface of a phospholipid membrane found that hydronium ions were bound near the lipid polar groups (phosphate and/or carbonyl groups) (*68*). The type 1 residues occupy a large fraction of the surface of the transmembrane protein that is exposed to the bulk solution. Considering the chemical similarity of lipid polar groups and the side chains of type 1 residues, they can form a continuous proton-conducing plane along the membrane surface (see Figure 8b). Here type 1 residues are needed to extend this plane to the proton entrance of the protein. However, without the type 2 residues, proton uptaking will be inefficient since the protons can also pass the entry point and continue the diffusion on the other side of the plane. Because of the relatively larger pK_a of the type 2 residues, the appearance

of type 2 residues forms a “proton sink” at the proton entrance (see Figure 8b). In this case, these two types of residues work together in a “ditch and sink” manner.

(a)



(b)



(c)

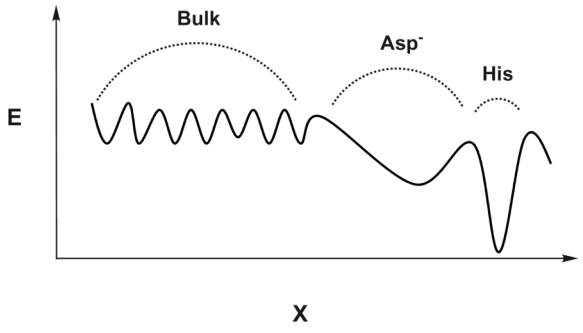


Figure 8: Possible functions of the co-appearing type 1 and type 2 residues according to (a) the "delocalized" theory (the "capture and store" mechanism) and (b) the "localized" theory (the "ditch and sink" mechanism). (a) illustrates the scenario, in which the proton is attracted by the type1 residues with their large capture radii (pink circle) and later transferred to or stabilized by the type2 residues (red circle). In (b), the cartoon shows the scenarios in which the proton diffuses along the membrane surface. The capture mechanism proceeds similar as to (a). (c) A qualitative energy surface of the two proton binding mechanism shown in (a) and (b). Here, X (also marked in (a)) denotes the relative position of the proton.

7.5 Conclusion

Q-HOP MD simulations revealed that 4-methylimidazole has a completely different protonation equilibrium compared to acetic acid in aqueous solution and with nearby proton accepting groups. Due to its relatively high pK_a $4MIH^+$ has a high tendency to keep a proton once it is bound. A close titratable group with lower pK_a only has few chances to snatch the proton from $4MIH^+$ that are driven by environmental fluctuations. On the other hand, 4MI has a relatively small proton capture radius, making it very hard to attract protons from long distances. Protonated acetic acid can easily share the proton with close titratable groups even if the acceptor group has a low pK_a. Moreover, AC^- has a large proton capture radius (about 5 Å), making it a perfect proton "capturer". Hydrogen bond chains involving the amino acid analogs, histidine and aspartic acid, are frequently found along proton transfer pathways in biomolecules. We suggest that the findings of this study on the model compounds 4MI and ACH are of relevance to biological proton transfer and this will be addressed in future work.

References

1. Iwata, S., Ostermeier, C., Ludwig, B., and Michel, H. (1995) Structure at 2.8-Angstrom Resolution of Cytochrome-C-Oxidase from Paracoccus-Denitrificans, *Nature 376*, 660-669.
2. Michel, H. (1999) Cytochrome c oxidase: Catalytic cycle and mechanisms of proton pumping-A discussion, *Biochemistry 38*, 15129-15140.
3. Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S., and Jap, B. K. (1998) Complete Structure of the 11-Subunit Bovine Mitochondrial Cytochrome bc1 Complex, *Science 281*, 64-71.
4. Xia, D., Yu, C.-A., Kim, H., Xia, J.-Z., Kachurin, A. M., Zhang, L., Yu, L., and Deisenhofer, J. (1997) Crystal Structure of the Cytochrome bc1 Complex from Bovine Heart Mitochondria, *Science 277*, 60-66.
5. Jang, S. S., Lin, S. T., Cagin, T., Molinero, V., and Goddard, W. A. (2005) Nanophase segregation and water dynamics in the dendrion diblock copolymer formed from the Frechet polyaryl ethereal dendrimer and linear PTFE, *J. Phys. Chem. B 109*, 10154-10167.
6. Marx, D. (2006) Proton Transfer 200 Years after von Grotthuss: Insights from Ab Initio Simulations, *ChemPhysChem 7*, 1848-1870.
7. Swanson, J. M. J., Maupin, C. M., Chen, H., Petersen, M. K., Xu, J., Wu, Y., and Voth, G. A. (2007) Proton Solvation and Transport in Aqueous and Biomolecular Systems: Insights from Computer Simulations, *J. Phys. Chem. B 111*, 4300-4314.
8. Olkhova, E., Helms, V., and Michel, H. (2005) Titration behavior of residues at the entrance of the D-pathway of cytochrome c oxidase from Paracoccus denitrificans investigated by continuum electrostatic calculations, *Biophys. J. 89*, 2324-2331.
9. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1997) On the quantum nature of the shared proton in hydrogen bonds, *Science 275*, 817-820.
10. Tuckerman, M. E., Laasonen, K., Sprik, M., and Parrinello, M. (1995) Ab Initio Molecular Dynamics Simulation of the Solvation and Transport of H3O+ and OH- Ions in Water, *J. Phys. Chem. 99*, 5749-5752.
11. Tuckerman, M. E., Laasonen, K., Sprik, M., and Parrinello, M. (1995) Ab initio molecular dynamics simulation of the solvation and transport of hydronium and hydroxyl ions in water, *J. Chem. Phys. 103*, 150-161.
12. Marx, D., Tuckerman, M. E., Hutter, J., and Parrinello, M. (1999) The nature of the hydrated excess proton in water, *Nature 397*, 601-604.
13. Riccardi, D., Konig, P., Prat-Resina, X., Yu, H. B., Elstner, M., Frauenheim, T., and Cui, Q. (2006) "Proton holes" in long-range proton transfer reactions in solution and enzymes: A theoretical analysis, *J. Am. Chem. Soc. 128*, 16302-16311.
14. Schmitt, U. W., and Voth, G. A. (1998) Multistate empirical valence bond model for proton transport in water, *J. Phys. Chem. B 102*, 5547-5551.
15. Vuilleumier, R., and Borgis, D. (1999) Transport and spectroscopy of the hydrated proton: A molecular dynamics study, *J. Chem. Phys. 111*, 4251-4266.
16. Vuilleumier, R., and Borgis, D. (1997) Molecular dynamics of an excess proton in water using a non-additive valence bond force field, *J.Mol. Struct. 437*, 555-565.
17. Vuilleumier, R., and Borgis, D. (1998) Quantum dynamics of an excess proton in water using an extended empirical valence-bond Hamiltonian, *J. Phys. Chem. B 102*, 4261-4264.
18. Schmitt, U. W., and Voth, G. A. (1999) The computer simulation of proton transport in water, *J. Chem. Phys. 111*, 9361-9381.
19. Day, T. J. F., Schmitt, U. W., and Voth, G. A. (2000) The mechanism of hydrated proton transport in water, *J. Am. Chem. Soc. 122*, 12027-12028.
20. Lobaugh, J., and Voth, G. A. (1996) The quantum dynamics of an excess proton in water, *J. Chem. Phys. 104*, 2056-2069.
21. Aqvist, J., and Warshel, A. (1993) Simulation of Enzyme-Reactions Using Valence-Bond Force-Fields and Other Hybrid Quantum-Classical Approaches, *Chem. Rev. 93*, 2523-2544.

22. Borgis, D., and Hynes, J. T. (1993) Dynamic Theory of Proton Tunneling Transfer Rates in Solution - General Formulation, *Chem. Phys. 170*, 315-346.
23. Borgis, D., and Hynes, J. T. (1996) Curve crossing formulation for proton transfer reactions in solution, *J. Phys. Chem. 100*, 1118-1128.
24. Staib, A., Borgis, D., and Hynes, J. T. (1995) Proton-Transfer in Hydrogen-Bonded Acid-Base Complexes in Polar-Solvents, *J. Chem. Phys. 102*, 2487-2505.
25. Marx, D., and Parrinello, M. (1996) Ab initio path integral molecular dynamics: Basic ideas, *J. Chem. Phys. 104*, 4077-4082.
26. Tuckerman, M. E., Marx, D., Klein, M. L., and Parrinello, M. (1996) Efficient and general algorithms for path integral Car-Parrinello molecular dynamics, *J. Chem. Phys. 104*, 5579-5588.
27. Park, J. M., Laio, A., Iannuzzi, M., and Parrinello, M. (2006) Dissociation mechanism of acetic acid in water, *J. Am. Chem. Soc. 128*, 11318-11319.
28. Ivanov, I., and Klein, M. L. (2002) Deprotonation of a histidine residue in aqueous solution using constrained ab initio molecular dynamics, *J. Am. Chem. Soc. 124*, 13380-13381.
29. Ivanov, I., Chen, B., Raugei, S., and Klein, M. L. (2006) Relative pK_a Values from First-Principles Molecular Dynamics: The Case of Histidine Deprotonation, *J. Phys. Chem. B 110*, 6365-6371.
30. Maupin, C. M., Wong, K. F., Soudackov, A. V., Kim, S., and Voth, G. A. (2006) A multistate empirical valence bond description of protonatable amino acids, *J Phys. Chem. A 110*, 631-639.
31. Pangali, C., Rao, M., and Berne, B. J. (1979) Monte-Carlo Simulation of the Hydrophobic Interaction, *J. Chem. Phys. 71*, 2975-2981.
32. Patey, G. N., and Valleau, J. P. (1975) Monte-Carlo Method for Obtaining Interionic Potential of Mean Force in Ionic Solution, *J. Chem. Phys. 63*, 2334-2339.
33. Chen, H., Wu, Y., and Voth, G. A. (2006) Origins of Proton Transport Behavior from Selectivity Domain Mutations of the Aquaporin-1 Channel, *Biophys. J. 90*, L73-75.
34. Bondar, A. N., Elstner, M., Suhai, S., Smith, J. C., and Fischer, S. (2004) Mechanism of primary proton transfer in bacteriorhodopsin, *Structure 12*, 1281-1288.
35. Mathias, G., and Marx, D. (2007) Structures and spectral signatures of protonated water networks in bacteriorhodopsin, *Proc. Natl. Acad. Sci. U. S. A. 104*, 6980-6985.
36. Lill, M. A., and Helms, V. (2001) Reaction rates for proton transfer over small barriers and connection to transition state theory, *J. Chem. Phys. 115*, 7985-7992.
37. Lill, M. A., and Helms, V. (2001) Molecular dynamics simulation of proton transport with quantum mechanically derived proton hopping rates (Q-HOP MD), *J. Chem. Phys. 115*, 7993-8005.
38. Lill, M. A., and Helms, V. (2001) Compact parameter set for fast estimation of proton transfer rates, *J. Chem. Phys. 114*, 1125-1132.
39. Lill, M. A., Hutter, M. C., and Helms, V. (2000) Accounting for environmental effects in ab initio calculations of proton transfer barriers, *J. Phys. Chem. A 104*, 8283-8289.
40. Lill, M. A., and Helms, V. (2002) Proton shuttle in green fluorescent protein studied by dynamic simulations, *Proc. Natl. Acad. Sci. U. S. A. 99*, 2778-2781.
41. de Groot, B. L., Frigato, T., Helms, V., and Grubmuller, H. (2003) The mechanism of proton exclusion in the aquaporin-1 water channel, *J. Mol. Biol. 333*, 279-293.
42. Gu, W., Frigato, T., Straatsma, T. P., and Helms, V. (2007) Dynamic Protonation Equilibrium of Solvated Acetic Acid, *Angew. Chem. Int. Ed. 46*, 2939-2943.
43. Hanggi, P., Talkner, P., and Borkovec, M. (1990) Reaction Rate Theory: Fifty Years After Krame, *Rev. Mod. Phys 62*, 251-342.
44. Herzog, E., Frigato, T., Helms, V., and Lancaster, C. R. D. (2006) Energy Barriers of Proton Transfer Reactions between Amino Acid Side Chain Analogs and Water from ab initio Calculations, *J. Comput. Chem. 27*, 1534-1547.
45. Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The Missing Term in Effective Pair Potentials, *J. Phys. Chem. 91*, 6269-6271.
46. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995), *J. Am. Chem. Soc. 118*, 2309-2309.
47. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A Smooth Particle Mesh Ewald Method, *J. Chem. Phys. 103*, 8577-8593.

48. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 3684-3690.
49. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes, *J. Comput. Phys. 23*, 327-341.
50. Jiang, F., McCracken, J., and Peisach, J. (1990) Nuclear quadrupole interactions in copper(II)-diethylenetriamine-substituted imidazole complexes and in copper(II) proteins, *J. Am. Chem. Soc. 112*, 9035-9044.
51. Michel, H. (1998) The mechanism of proton pumping by cytochrome c oxidase, *Proc. Natl. Acad. Sci. U. S. A. 95*, 12819-12824.
52. Jormakka, M., Trnroth, S., Byrne, B., and Iwata, S. (2002) Molecular Basis of Proton Motive Force Generation: Structure of Formate Dehydrogenase-N, *Science 295*, 1863-1868.
53. Feher, G., Allen, J. P., Okamura, M. Y., and Rees, D. C. (1989) Structure and function of bacterial photosynthetic reaction centres, *Nature 339*, 111-116.
54. Hunte, C., Screpanti, E., Venturi, M., Rimon, A., Padan, E., and Michel, H. (2005) Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH, *Nature 435*, 1197-1202.
55. Shi, N., Ye, S., Alam, A., Chen, L., and Jiang, Y. (2006) Atomic structure of a Na+- and K+-conducting channel, *Nature 440*, 570-574.
56. Gonen, T., Cheng, Y., Sliz, P., Hiroaki, Y., Fujiyoshi, Y., Harrison, S. C., and Walz, T. (2005) Lipid-protein interactions in double-layered two-dimensional AQP0 crystals, *Nature 438*, 633-638.
57. Ferguson-Miller, S., and Babcock, G. T. (1996) Heme/copper terminal oxidases, *Chem. Rev. 96*, 2889-2907.
58. Berg, B. L., Li, J., Heider, J., and Stewart, V. (1991) Nitrate-inducible Formate Dehydrogenase in Escherichia coli K-12, *J. Biol. Chem. 266*, 22380-22385.
59. Ermler, U., Fritzsch, G., Buchanan, S. K., and Michel, H. (1994) Structure of the photosynthetic reaction centre from Rhodobacter sphaeroides at 2.65 A resolution: cofactors and protein-cofactor interactions, *Structure 2*, 925-936.
60. Paddock, M. L., Feher, G., and Okamura, M. Y. (2000) Identification of the proton pathway in bacterial reaction centers: Replacement of Asp-M17 and Asp-L210 with Asn reduces the proton transfer rate in the presence of Cd2+, *Proc. Natl. Acad. Sci. U. S. A. 97*, 1548-1553.
61. Paddock, M. L., Graige, M. S., Feher, G., and Okamura, M. Y. (1999) Identification of the proton pathway in bacterial reaction centers: Inhibition of proton transfer by binding of Zn2+ or Cd2+, *Proc. Natl. Acad. Sci. U. S. A. 96*, 6183-6188.
62. Stowell, M. H. B., McPhillips, T. M., Rees, D. C., Soltis, S. M., Abresch, E., and Feher, G. (1997) Light-Induced Structural Changes in Photosynthetic Reaction Center: Implications for Mechanism of Electron-Proton Transfer, *Science 276*, 812-816.
63. Padan, E., Tzubery, T., Herz, K., Kozachkov, L., Rimon, A., and Galili, L. (2004) NhaA of Escherichia coli, as a model of a pH-regulated Na+/H+ antiporter, *Biochim. Biophys. Acta-Bioenergetics 1658*, 2-13.
64. Olkhova, E., Hunte, C., Screpanti, E., Padan, E., and Michel, H. (2006) Multiconformation continuum electrostatics analysis of the NhaA Na+/H+ antiporter of Escherichia coli with functional implications, *Proc. Natl. Acad. Sci. U. S. A. 103*, 2629-2634.
65. Antonenko, Y. N., Kovbasnjuk, O. N., and Yaguzhinsky, L. S. (1993) Evidence in favor of the existence of a kinetic barrier for proton transfer from a surface of bilayer phospholipid membrane to bulk water, *Biochim. Biophys. Acta 1150*, 45-50.
66. Heberle, J., Riesle, J., Thiedemann, G., Oesterhelt, D., and Dencher, N. A. (1994) Proton migration along the membrane surface and retarded surface to bulk transfer, *Nature 370*, 379-382.
67. Prats, M., Teissie, J., and Tocanne, J.-F. (1986) Lateral proton conduction at lipid-water interfaces and its implications for the chemiosmotic-coupling hypothesis, *Nature 322*, 756-758.
68. Smondyrev, A. M., and Voth, G. A. (2002) Molecular dynamics simulation of proton transport near the surface of a phospholipid membrane, *Biophys. J. 82*, 1460-1468.

Chapter 8

Conclusion and Outlook

In this thesis, several biological/chemical systems were studied using standard and variant molecular dynamics simulation techniques.

The first 4 chapters (chapters 2–5) present studies of the interaction of proline-rich peptides with their adaptor domains (mainly the GYF domain). The recognition of proline-rich sequences plays an important role for the assembly of multi-protein complexes during the course of eukaryotic signal transduction and is mediated by a set of protein folds that share characteristic features. We studied the solvent conformations of the wild-type (SHRPPPPGHRV) and mutant (SHRPPPPWHRV, and SHRPPPPGMRV) polyproline peptides that bind to the GYF domain using molecular modeling and MD simulations. The simulations of the wild-type peptides were carried out using different starting structures (complex conformation and extended conformation) and under different conditions (room temperature and high temperature). We found that in all simulations, the peptides formed PPII helix conformations even in the absence of the GYF domain. These results agree well with recent experimental and theoretical studies on polypeptides with or without prolines and indicate that the formation of a PPII helix of the peptide is not induced by the binding processes alone. They also reveal the first step of the binding process: the preformation of the PPII helix, i.e. the enthalpy and entropy cost of this step does not contribute to the binding affinity.

MD simulations of the complex structures were performed on the wild-type complex as well as on the modeled and the docked mutant complexes (G8W, H9M, G8R, G8Y, and G8K). Based on our previous knowledge from NMR experimental studies of the GYF domain-ligand interaction and the simulations of the wild-type and mutated complexes, we modeled the general binding mode of polyproline peptides to the GYF domain. The hydrophobic interactions between the peptide residues Pro6 and Pro7, and binding pocket as well as the electrostatic attractions between the peptide residues Arg3 and Arg10, and the domain residues Glu31 and Glu9 play crucial roles in the binding. These important interactions are proved by our recent MD studies on the whole binding processes, where the individual roles of each interaction pair are identified (not shown in this thesis, work together with Mazen Ahmad). By peptide docking and subsequent MD simulations of the G8X mutants, we identified an alternative binding mode, where a shift in register for the interacting prolines was observed. These results agree qualitatively well with NMR chemical shift mapping experiments and indicate dynamic processes to be important for proline-rich sequence

recognition. We suggested that such gliding motions along long proline-rich sequences decrease the entropic penalty of binding while still keeping a certain degree of specificity. This may be a possible explanation of the involvement of the interaction partners in the so-called “transient interactions”, where ligands and receptors bind and unbind in a relatively “fast” manner.

We also studied the binding of linear peptide motifs to Cyclophilin A using peptide docking and MD simulation methods. Our experimental collaborators identified the linear sequence recognition code for CypA and the consensus motif FGPXLp using substitution analysis (phage display). Our modeled complex structure agrees very well with the results form phage display experiments and gives an explanation of the specific binding motif from structural and interaction points of view.

Since the solvation effects are crucial in almost every process in molecular biology, and in order to account for this effect in the future development of residue-scale screening methods for biomolecular interactions, we computed the solvation free energies of peptides of various lengths using the multi-configuration thermodynamic integration (MCTI) with separation-shifted potential scaling method. Similar calculations were also carried out using the Generalized-Born Surface Area (GBSA) solvation model for static helix geometries. In this study we found that for 5 or more residues the results obtained from both methods are in quite good agreement. This observation gives strong support for our strategy of computing ΔG_{hydr} for peptides up to 9 residues from MCTI calculations. However, MCTI and GBSA still show sizable differences for short helices where MCTI should be quite accurate. This indicates that it is important to consider molecular details of backbone hydration. Non-additivity of the solvation free energy is found for peptides shorter than 5 residues, while additivity appears fulfilled for helices longer than 10 residues. The non-additivity of the solvation free energy makes the design of simplified models a challenging task, since in many of such models peptides or proteins are composed of residue-beads and the interactions are modeled additively. In this collaborative work with Dr. René Staritzbichler, I was involved in the MCTI and GBSA calculations of the peptides, summarized the results and wrote the manuscript.

To describe the complete binding process and mechanism of polyproline peptides binding to their recognition domains, further simulations are certainly needed. Possible strategies may be to combine Brownian dynamics simulations and MD simulations, or MD simulation with implicit solvent models. The newly developed computer hardware and simulation software provide much stronger computing power than available at the time when the above-mentioned simulations were carried out. Current computational resources make it also possible to simulation the whole binding process using MD simulations, e.g., on the recognition domain with random placed peptides to discover the diffusion properties and the binding steps. For large scale screening of peptide sequences that can bind to the GYF domain, more efficient methods, e.g. residue-scale model, peptide docking, or statistical learning, are certainly needed.

Chapters 6 and 7 present our studies on the dynamic protonation equilibria of amino acid side chain analogs in aqueous solution using the Q-HOP MD methodology. Protonation equilibria are essential in many biological and chemical processes. In biological systems, especially in membrane proton pumps, the proton

transfer mechanisms are rather complicated and the principles behind these mechanisms are best revealed by transferring knowledge obtained on well-understood model systems. The Q-HOP method was developed by Lill and Helms based on classical MD simulation accounting for stochastic proton transfer events. In this thesis, we reported a better treatment of the environmental contribution of the Q-HOP model by optimizing a separate set of q_{12}^{env} charges to reproduce the results from QM/MM calculations.

Using the Q-HOP MD simulation technique and the new environmental charges, we studied the protonation equilibria of acetic acid (AC^-/ACH, side chain analog of aspartic acid) and 4-methylimidazole ($4MIH^+$/4MI, side chain analog of histidine) in aqueous solution and with nearby proton accepting groups. In the simulation of solvated acetic acid, two different regimes of proton transfer were observed: where the proton either frequently swaps between acetic acid and nearby water or freely diffuses in the simulation box until it is captured again by acetic acid. We calculated the pK_a of acetic acid based on the relative population of protonated and deprotonated states. The diffusion coefficient of the excess proton was also computed from the average mean squared displacement in the simulation. Both calculated values agree well with the experimental measurements. This is the first work where the dynamic protonation equilibrium between an amino acid side chain analogue and bulk water as well as the diffusion properties of the excess proton were successfully reproduced through unbiased computer simulations. It serves as a first exploration of a dynamic protonation equilibrium by computer simulation as well as a control of the Q-HOP model.

In the study of 4-methylimidazole in aqueous solution and with nearby proton accepting groups, a qualitatively different protonation behavior of 4-methylimidazole compared to that of acetic acid was found: Due to its relatively high pK_a $4MIH^+$ has a high tendency to keep a proton once it is bound. A close titratable group with lower pK_a only has few chances to snatch the proton from $4MIH^+$ that are driven by environmental fluctuations. On the other hand, 4MI has a relatively small proton capture radius, making it very hard to attract protons from long distances. Protonated acetic acid can easily share the proton with close titratable groups even if the acceptor group has a low pK_a. Moreover, AC^- has a large proton capture radius (about 5 Å), making it a perfect proton "capturer".

Hydrogen bond chains involving the amino acid analogs histidine and aspartic acid are frequently found along proton transfer pathways in biomolecules. Therefore, we suggest that the findings of this study on the model compounds 4MI and ACH are of relevance to biological proton transfer. In future work, a complete protein and lipid bilayer in solvent environment will be included. In this way, the functions of aspartic acid/glutamic acid and histidine at the proton entry of membrane proton pumps like cytochrome *c* oxidase will be studied. Further problems such as proton transfer pathways in those pumps can be studied as well.

List of Publications

1. **Wei Gu**, Jiang Zhu and Haiyan Liu
   Different Protonation States of The Bacillus Cereus Binuclear Zinc Metallo-β-Lactamase Active Site Studied by Combined Quantum Mechanical and Molecular Mechanical Simulations
   J. Theor. Comp. Chem., **1**, 69-80 (2002)

2. **Wei Gu**, Tingting Wang, Jiang Zhu, Yunyu Shi and Haiyan Liu
   Molecular dynamics simulation of the unfolding of the human prion protein domain under low pH and high temperature conditions
   Biophys. Chem., **104**, 79-94 (2003)

3. **Wei Gu**, Sahand J. Rahi and Volkhard Helms
   Solvation Free Energies and Transfer Free Energies for Amino Acids from Hydrophobic Solution to Water Solution from a Very Simple Residue Model
   J. Phys. Chem. B, **108**, 5806-5814 (2004)

4. Saurabh K. Shakya, **Wei Gu**, Volkhard Helms
   Molecular Dynamics Simulation of Truncated Bovine Adrenodoxin
   Biopolymers, **78**, 9-20 (2005)

5. **Wei Gu**, Michael Kofler, Iris Antes, Christian Freund and Volkhard Helms
   Alternative Binding Modes of Polyproline Peptides Binding to the GYF Domain
   Biochemistry, **44**, 6404-6415 (2005)

6. Kirill Piotukh, **Wei Gu**, Michael Kofler, Volkhard Helms and Christian Freund
   Cyclophilin A binds to linear peptide motifs containing a consensus that is present in many human proteins
   J. Biol. Chem., **280**, 23668-23674 (2005)

7. Rene Staritzbichler, **Wei Gu** and Volkhard Helms
   Are solvation free energies of homogeneous helical peptides additive?
   J. Phys. Chem. B, **109**, 19000-19007 (2005)

8. **Wei Gu** and Volkhard Helms
   Dynamical Binding of Proline-rich Peptides to their Recognition Domains
   Biochim. Biophys. Acta - Proteins and Proteomics, **1754**, 232-238 (2005)

9. **Wei Gu**, Tomaso Frigato, Tjerk P. Straatsma and Volkhard Helms
   Dynamic Protonation Equilibrium of Solvated Acetic Acid
   Angew. Chem. Int. Ed., **46**, 2939-2943 (2007)
   Dynamisches Protonierungsgleichgewicht der in Wasser geloesten Essigsaeure
   Angew. Chem., **119**, 2997-3001 (2007)

10. **Wei Gu** and Volkhard Helms
    Different Protonation Equilibria of 4-Methylimidazole and Acetic Acid
    ChemPhysChem, **8**, 2445-2451 (2007)