Dissertation
submitted to the
Combined Faculties for Natural Sciences and Mathematics
of the Ruperto Carola University of
Heidelberg, Germany,
for the degree of
Doctor of Natural Sciences

presented by
Diplom–Biochemiker Raik Grünberg
born in Riesa

oral examination:

Proteins on the edge

Transitions of structure ensembles in protein unfolding and protein-protein binding

Referees: Prof. Dr. Jeremy C. Smith
Dr. Michael Nilges

Contents

Zusammenfassung

Proteine sind ständig in Bewegung. Diese Beweglichkeit speist sich aus dem komplexen Wechselspiel tausender Atome. Die experimentelle Struktur – mit ihren exakten Koordinaten für jedes Atom – ist also in Wirklichkeit nur der Mittelwert einer vielfältigen Mischung von Konformationen. Bewegung ist oft das Bindeglied zwischen Proteinstruktur und biologischer Funktion, erweist sich aber gleichzeitig als einer der am wenigsten verstandenen Aspekte der Strukturbiologie. In der vorliegenden Arbeit untersuche ich die Dynamik von Proteinen "auf der Kippe", also im Grenzbereich zwischen zwei Zuständen. Wie es scheint, kommt die in der Struktur verborgene Vielfalt gerade dann zum Tragen, wenn sich das komplexe Molekül im Ungleichgewicht, im Übergang oder, anders ausgedrückt, in biologischer Aktion befindet.

Netzwerke aus in vielfacher Kopie aneinandergereihten Spektrindomänen verleihen der Membran von Erythrozyten bemerkenswerte Elastizität. Die Beweglichkeit der Domäne hinterlässt deutliche Spuren in mechanischen Entfaltungsexperimenten an einzelnen Molekülen. Wie Simulationen zeigen, entscheiden zufällige Fluktuationen, wie lange sich die Spektrindomäne mechanischer Belastung widersetzt und ob nicht-native Strukturen die vollständige Entfaltung aufhalten. Diese Unschärfe der einzelnen Glieder, gemittelt über die gesamte Kette, bedingt vermutlich eine gleichmäßige Rückstellkraft über einen sehr weiten Dehnungsbereich. Experimente an gezielt veränderten Spektrindomänen unterstützen dieses Bild. Die Elastizität roter Blutzellen beruht also vielleicht auch auf der chaotischen Bewegung einzelner Proteinabschnitte.

Die meisten Proteine agieren nicht allein, sondern finden sich eingebettet in ein dichtes Netz von Wechselwirkungen. Fluktuationen der Struktur haben offenbar beträchtlichen Einfluss sowohl auf die Stabilität von Proteinkomplexen als auch auf die Geschwindigkeit ihrer Bildung. Das komplexe Zusammenspiel von Proteindynamik und der Wechselwirkung zwischen Proteinen entzieht sich aber bisher weitestgehend unserem Verständnis. Ich vergleiche die Dynamik von 17 Proteinkomplexen und den daran beteiligten Partnern. Wie die umfangreichen Simulationen enthüllen, sind freie Bindungsstellen deutlich flexibler als die restliche Oberfläche des Proteins. Entgegen der üblichen Annahme wird aber die allgemei-

ne Beweglichkeit der Proteine im Komplex nicht grundsätzlich eingeschränkt. Die Bindung kann sowohl mit dem Verlust als auch mit dem Gewinn von konformeller Entropie einhergehen. Auch die Vorstellungen vom Erkennungsvorgang selbst ziehen die Flexibilität von Proteinen bisher kaum in Betracht. Ich verknüpfe die Simulationen mit einer systematischen Untersuchung der Passgenauigkeit zwischen verschiedenen Konformationen der beiden Bindungspartner. Erkennung erfordert oft spezifische Varianten der freien Struktur. Mein erweitertes Modell für den Mechanismus der Proteinbindung trägt dem Rechnung und erscheint besser vereinbar mit theoretischen und experimentellen Daten.

Summary

Proteins move. Their incessant fluctuations are governed by a complex interplay between thousands of atoms. Experimental structures, providing exact coordinates for every atom, hence only represent the average of a diverse ensemble of interchanging conformations. Molecular motion is often the barely understood link between structure and biological function. The present work examines two different processes that put proteins on the edge of moving from one global state to another. At the moment of transition, perturbation or, indeed, biological action, benign structure fluctuations can, it seems, turn into major forces.

Chains of spectrin repeats apparently rely on structure flexibility to achieve a smooth response to external force. Single molecule atomic force microscopy experiments on this domain, in accord with simulations, showed clear traces of structure fluctuation. On the verge of disruption, thermal fluctuations decide how much extension a spectrin repeat tolerates and whether or not unfolding is blocked by intermediate non-native structures. This picture was supported by experiments and simulations on mutated repeats. The elasticity of the membrane skeleton and, for example, red blood cells, may thus to some extent depend on chaotic motions within single protein domains.

Structure fluctuations also affect the process of protein-protein interaction, but the interplay of protein flexibility and recognition remains far from understood. I performed and compared molecular dynamics simulations on 17 protein complexes as well as their free components. Free interaction patches turned out more flexible than the remaining protein surface. However, contrary to common sense, binding does not generally restrict protein flexibility and conformational entropy may be lost but also gained in the process. Current models of recognition do not account for overall protein flexibility or make assumptions that are incompatible with kinetic observations. I combined the simulation data with systematic docking calculations and derived a new model for this process. Often, only subsets of the two free structure ensembles were mutually compatible. A conformer selection step may thus impede the rate of recognition. Protein fluctuations seem to be actively involved in the binding reaction and influence or even control the speed of recognition as well as the stability of the complex.

Introduction

Proteins are the nanoscopic workhorses of life. Cells rely on proteins to break down and build chemical compounds, to process information, and to withstand or generate mechanical forces. These and a variety of other tasks are performed by molecules that are chemically very much alike. All proteins are ultimately made up from the same 20 amino acids[1] which are linked into polypeptide chains. Depending on the order (sequence) of amino acids, polypeptide chains fold into diverse three-dimensional structures. The individual spatial arrangement of their atoms confers vastly different properties and functions to, from a pure chemist's point of view, very similar molecules. The first three-dimensional structure of a protein, myoglobin, was determined in 1959 by John Kendrew et al. (1960). Since then, structure was considered the missing link between the amino acid sequence and the biological function of a protein. Max Perutz (Perutz 1970; Perutz et al. 1998) demonstrated in decades of work on hemoglobin how atomic structure gives rise to physiological activity as well as disease.

Inspired by this work, a new scientific discipline – structural biology – took on the task to explain life in terms of atomic structure. Yet, increasingly it became clear that proteins don't actually have a single structure. To some extend, this was known from the outset, since already hemoglobin had been crystallized in two different structural variants – one with and one without oxygen bound. Such a "switch" between conformations - bound and unbound, active and inactive or (in case of prions) even benign and infective – is hallmark of many proteins and crucial for enzymatic activity, signaling cascades, and various other aspects of cell

[1] some procaryotic and eucaryotic genes code for a rare 21^{st} amino acid – Selenocysteine

biology. Not surprisingly, the transition between distinct conformational states received much attention.

However, even this description turned out to be oversimplified. The distinct ”conformational states” themselves are far from static. Proteins move. Experimental protein structures are only the average of a dense ensemble of diverse and rapidly interchanging subconformations (Frauenfelder et al. 1991). The stability and ongoing movement of a protein molecule is governed by the complex mixture of forces between its usually many thousand atoms which are further influenced by many thousand solvent molecules. Transitions from ”one” global conformational state to ”another” thus require the collective action of many interacting particles. The collective behavior of such many-body systems, especially outside equilibrium, is notoriously difficult to capture with physical theories, and one must therefore resort to computationally expensive simulations. The dynamics of protein structure thus rank among the most challenging problems to be understood with current physical methods, next to atmospheric circulation, cosmic processes, and the (entirely dispensable but probably better funded) design of nuclear weapons[2]. In fact, at the time of writing, of the two (officially) most powerful computers worldwide, one is dedicated to climate and earth quake simulations and the other to the simulation of protein folding.

The research on protein folding was perhaps the first field in structural biology that fully embraced the new ensemble view on protein structure. The transition from a disordered peptide chain to the folded protein had traditionally been described by one-dimensional energy profiles with exact intermediate states. Long debates about one or another folding intermediate were finally resolved by the new model of an energy landscape that funnels a large variety of conformations along many routes toward the native structure ensemble (Dill and Chan 1997). Beyond the folding process, the implications of protein flexibility are, nevertheless, still somewhat ignored in several areas of structural biology. An isolated structure is usually not enough to understand or predict the behavior of a protein.

[2]Another obvious problem worth studying would be how societies of millions of interacting people arrive at spending much of their resources on atomic or conventional bombs – of course there is hardly funding for that kind of research.

On the one hand, proteins work in the context of a cell where they are embedded in a web of mutual interactions and modifications. On the other hand, the flexibility of these complex macromolecules is without doubt crucial for many of their functions. Both aspects constitute a missing link between protein structure and function; both are touched in the following chapters.

In this work, I use established simulation techniques to study the dynamics of proteins that are on the edge of moving from one global state to another. In the first part (chapter 2) I analyze the response of a "mechanical" protein domain to external force. Structure diversity (or flexibility) leaves remarkable traces in experiments that monitor the forced unfolding of single spectrin repeats. In fact, these protein domains appear to utilize structure fluctuations at the atomic level to achieve macroscopic elasticity. In other words, they translate flexibility into biological function. The second part (chapter 3) examines the interplay of protein flexibility and protein-protein interaction. Instead of analyzing a single protein, this chapter follows a more comparative approach. I perform simulations on several complexes as well as their free binding partners and extract trends that may hold for protein-protein binding in general. Dynamic properties distinguish free interaction patches from the remaining protein surface. However, contrary to common sense, binding does not generally restrict protein flexibility and conformational entropy may be lost but also gained in the process. Current models of the recognition process do not account for overall protein flexibility or make assumptions that are incompatible with kinetic observations. I combine the simulation data with systematic docking calculations and derive a new model for this process. According to these results, protein fluctuations may be actively involved in the binding reaction and influence or even control the speed of recognition as well as the stability of the complex.

Chapter 2 is based on the joint publication with an experimental group (Altmann et al. 2002) but puts our results into a somewhat sharper focus. Chapter 3 recapitulates another recent publication (Grünberg et al. 2004) and is extended by yet unpublished data. The two parts are wrapped up by a concluding chapter. In the beginning, I provide a brief introduction to methods and models for the study of protein structure and dynamics.

Chapter 1

Structures and ensembles

1.1 Introduction

Proteins usually adopt a defined, but not static, three-dimensional structure. A complex interplay of compensating forces locks the chain (or chains) of bonded amino acids into a small section of the vast space of conformational possibilities. Protein atoms are held together by strong chemical bonds, many of which, however, are rotatible and hence leave large conformational freedom to the complex molecule. This freedom is restricted by long reaching electrostatic interactions (attractive or repulsive). It is further restricted by short range van der Waals (or dispersion) forces which stem from temporary dipole moments that are mutually induced in adjacent electronic shells. The system of protein and surrounding solvent then adopts the state with the maximum number of energetically still accessible degrees of freedom. Entropy of solvent and solute are thus another important component within the delicate balance of forces.

The following sections give a brief survey of the basic principles of, first, the experimental and, second, the computational methods that are commonly used for the study of protein structure and dynamics. The emphasis, especially of the second section, rests with methods that have been relevant to my work. The description is by no means exhaustive – neither in coverage nor in depth. The chapter ends with a short (and subjective) review of how ideas and models have been

evolving to add more and more dynamics to our picture of protein structure and function.

1.2 Experimental methods for the study of protein dynamics

1.2.1 Structure determination

Structural biology depends on structures. X-ray crystallography was the first and is until today the most frequently used method to obtain atomic structures of proteins and other macromolecules. Other than visible light, X-rays have a short enough wavelength (about 1 Å) to probe molecules at atomic detail. However, the direct "X-ray microscopy" of a single molecule is not yet possible. The signal would be too weak and the radiation would furthermore very quickly destroy the object (planned free electron X-ray lasers may resolve both problems, see section 1.2.2). The necessary information is hence collected from a large number of molecules "in phase". In a crystal of molecules, X-rays are diffracted at the planes that are formed by the repetitively arranged atoms. Monochromatic beams interfere with each other if they are reflected from parallel planes. The interference is destructive unless the Bragg condition is fulfilled:

$$n\lambda = 2d\sin\vartheta \tag{1.1}$$

If the staple of parallel planes, spaced evenly at distance d, is hit at Bragg angle ϑ, the beam travels exactly one or a multiple of its wave lengths λ between two planes and is hence amplified by positive interference. X-ray crystallography hence roughly consists of (1) preparing a crystal of the protein and (2) recording the diffraction pattern of an X-ray beam passing through this crystal at different angles. The structure information is encoded by the intensities measured at the various reflective angles. However, the pattern of intensities lacks any information about the phase of the interfering waves. In a third step, this phase information is reconstructed from homology models, or by using scattering from heavy metal,

or selenium and sulfur atoms incorporated in the crystal. X-rays interact with the molecule's electrons and the diffraction pattern (plus phase information) can hence be transformed into a three-dimensional electron density map. The covalent protein structure is then modeled into this map, but, due to their low electron densities, typically lacks the hydrogen positions. Protein expression and, especially, crystallization usually constitute the main bottleneck of the procedure.

The second important method for structure determination examines molecules directly in solution, without the need of protein crystals. Nuclear Magnetic Resonance (NMR) spectroscopy measures the magnetization of atomic nuclei. Atomic nuclei with a spin I different from zero have a magnetic moment $\mu = \gamma I$ that depends on their gyromagnetic constant γ. They have $2I + 1$ energetically distinct possibilities to align to a magnetic field. For example, ^{1}H, ^{13}C or ^{15}N nuclei with $I = \frac{1}{2}$ can adopt two "orientations" that differ in their magnetic quantum number m_I. This gives them two possible energy values E_I when exposed to an external magnetic field B_0:

$$E_I = -\mu B_0 = \gamma m_I B_0 \,, \quad m_I = \frac{1}{2}, -\frac{1}{2} \tag{1.2}$$

Transitions between these quantum states can be incited by an electromagnetic pulse of the right frequency ν to bridge the energy gap ΔE:

$$h\nu = \Delta E = \gamma B_0 \tag{1.3}$$

However, the resonant frequency not only depends on the external field B_0 (and the type of element) but is also influenced by the atom's local environment. This leads to a *chemical shift* of the atom's resonant frequency and allows the spectroscopist to distinguish, for example, the signals from the different hydrogen atoms of a protein. The magnetization of covalently attached neighbors can furthermore split an atom's signal into multiple resonances. The extent of this spin coupling depends on the angles of connecting bonds and thus yields first structural information. However, most important for structure determination is the *Nuclear Overhauser Effect* (NOE). The NOE is caused by the dipole – dipole coupling

between nuclear spins, which transfers spin polarization between different nuclei. The intensity of this NOE depends on the spatial distance of the two atoms involved ($\propto r^{-6}$). The classic approach of NMR structure determination hence requires first experiments that assign the observed frequencies ν to individual atoms of the molecule, followed by measurements that probe all of these atoms for NOEs with the others. The result is a set of distance restraints which is combined with *a-priori* knowledge of bond lengths, angles and atomic radii to obtain a structural model of the protein.

Compared to X-ray crystallography, NMR has the advantage to study proteins in solution and to circumvent the critical crystallization step. However, the spectrum of nuclear resonances becomes increasingly complex for larger molecules. Isotope labeling, multi-dimensional experiments and other techniques help alleviating this problem. Nevertheless, for full-fledged structure determination, NMR remains largely restricted to proteins below 40 kD mass. On the other hand, the method reveals a diverse set of additional information and is a versatile tool for the study of protein dynamics. This will be discussed in the following section.

Protein structures can also be determined by other, less routinely employed, techniques. Neutron diffraction, for example, operates on principles very similar to X-ray crystallography. In contrast to the latter, protein crystals are probed with neutron radiation that interacts with atomic nuclei, and the diffraction pattern hence yields atom density maps, rather than electron densities. The experiments can be carried out at physiological temperatures and give the position of hydrogen atoms (Engler et al. 2003). By contrast, X-ray diffraction often requires very low temperatures to limit radiation damage and hydrogen positions usually have to be ”guessed” by a modelling program. This prediction is, in detail, often inaccurate (Engler et al. 2003). However, among other technical issues, neutron diffraction requires large protein quantities (i.e. large crystals) of high stability and relies on expensive and rare neutron sources.

Several other experimental methods reveal partial structural information. The approximate shape of macromolecules can be determined by neutron or X-ray scattering in solution as well as by electron microscopy. Different spectroscopic techniques add further valuable data.

1.2.2 Dynamic properties

Standard X-ray crystallography reveals only very limited information about the dynamics of a protein. B-factors are meant to quantify the atom's mean square displacement from the given (average) coordinate. However, apart from thermal motion, they are also influenced by other factors, namely partial disordering, whole molecule or domain displacements, crystal defects, and inaccuracies of the model building (Petsko and Ringe 1984). Furthermore, B-factors contain no information about the time scale (or frequency) of the underlying vibrations.

Instead of irradiating it with monochromatic X-rays at different angles, protein crystals can also be probed by a polychromatic beam at one angle. This Laue diffraction could theoretically shorten data collection to the length of a single X-ray pulse (Zhong et al. 1999), about 100 ps on current synchrotron sources. Unfortunately, the intensity is not yet high enough and data have to be accumulated over several pulses or with longer exposure times (Zhong et al. 1999). Nevertheless, Laue diffraction in combination with sophisticated synchrotron sources allows for time resolved X-ray crystallography of down to nanosecond resolution (Schlichting and Chu 2000). Yet, the resulting "snapshots" still constitute an average over many billion molecules of the crystal. Rather than arbitrary nanosecond-scale motions, the technique therefore only reveals movements that are followed simultaneously by the whole ensemble. Such perfect synchronization cannot be achieved but it has been for example possible to observe intermediate states of enzymatic reactions (Schlichting and Chu 2000).

The arrival of free electron X-ray lasers may remove the obstacle of crystallization from X-ray crystallography. X-ray pulses from anticipated new sources should be bright enough to record diffraction data from single protein molecules and short enough to do so before the destruction of the object (Neutze et al. 2000; Miao et al. 2004). However, the method will probably still require averaging over many molecules and it remains to be seen whether it can provide data on protein dynamics.

Incoherent neutron scattering allows to measure atomic fluctuations, especially of hydrogen atoms, on time scales typically reaching up to 100 ps (Zac-

cai 2000). The method yields the mean displacement of ^{1}H atoms of a protein. Deuterium labeling can be used to focus on sub-sets of atoms. The results are not atom-specific but more accurate and direct then values extracted from X-ray crystallographic B-factors, and the method does not require crystalline samples.

In contrast to the techniques described so far, spectroscopic methods give access to both frequency and amplitude of certain molecular motions and thus constitute an important tool for the study of protein dynamics.

Moessbauer spectroscopy was instrumental for early works on protein flexibility (Keller and Debrunner 1980). Nuclei of several elements, for example ^{57}Fe, absorb and emit γ rays at very defined wave lengths as long as the atom is part of condensed matter, for example a protein crystal. Movements of the atom relative to the γ source shift this wave length due to the Doppler effect. Moessbauer spectroscopy measures fast vibrations of a particular atom very precisely, but is naturally restricted to proteins that contain such a sensitive atom at a defined position.

Meanwhile, NMR relaxation spectroscopy has emerged as a primary tool for the study of protein dynamics (reviewed by Bruschweiler (2003)). NMR relaxation experiments excite various of the protein's nuclear spins and then observe the decay of this magnetization over time. This spin relaxation is promoted by internal protein motions but also by the rotational tumbling of the whole protein. The relaxation falls into two different time ranges. Longitudinal relaxation times (T_1) are modulated by motions on a nanosecond and subnanosecond time scale; Transverse relaxation (T_2) sheds light onto microsecond to millisecond dynamics. T_1 relaxation data are typically transformed into order parameters that characterize the mobility of certain bond vectors and can also be related to conformational entropy. The peptide N-H bond is most accessible to this kind of measurements and the backbone dynamics of many proteins have been characterized on a per-residue basis. There are some caveats to these data and their analysis. The calculation of order parameters is based on the assumption that overall and local motions are separable and thus ignores the significant correlations that exist between the motions of different residues (Bruschweiler 2003). Moreover, as I will

show in section 3.3.6, backbone dynamics appear not generally representative for a protein's overall mobility, at least not during protein-protein binding. On the other hand, NMR relaxation experiments are increasingly extended to certain side chain bonds and the formalisms for their analysis benefit from ongoing development (Bruschweiler 2003).

In summary, different physical methods examine different time scales of protein motion. However, all of these methods reveal only indirect and partial information. There is yet no single molecule "real-time imaging" technique. Simulation techniques have thus become an important tool for connecting and interpreting the patchwork of experimental data.

1.2.3 Single molecule experiments

Experiments are usually performed on systems comprising billions or trillions of molecules and hence observe average properties of the whole ensemble. Molecular simulations, on the other hand, examine single or very few molecules. The comparison of simulation results with macroscopic and measurable quantities often turns out anything but straightforward. This gap is increasingly filled by a new class of experiments that detect and manipulate single molecules.

Two fluorescing dyes with overlapping absorption frequencies can act as fluorescence donor and acceptor, that is the excitation of the donor dye leads to the fluorescence of the acceptor. The efficiency of this fluorescence resonance energy transfer (FRET) depends on the distance between donor and acceptor and can hence serve as a "molecular ruler" for distances between 10 and 75 Å. Modern detectors and optics measure the fluorescence transfer between single dye molecules. It thus becomes possible to follow the distance of two dyes that are attached to a single protein and, for example, to observe large-scale movements of subunits or domains (reviewed by Weiss (1999)).

Another set of techniques goes one step further and allows not only to observe but also to manipulate single molecules. Atomic force microscopes (AFM) (Binnig et al. 1986) as well as optical or magnetic tweezers (Ashkin 1987; Smith et al. 1992) capture and move single proteins or micrometer-sized particles with

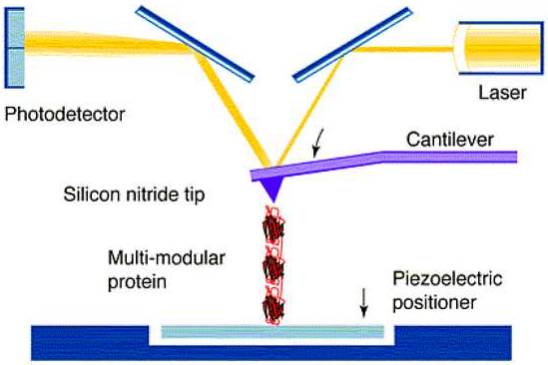

Figure 1.1: Atomic force microscopy. [reproduced with kind permission from Fisher et al. (1999)]

single proteins attached. In the case of optical tweezers, the particle is dragged to the waist of a focused laser beam. Deviations from this position correspond to picoNewton forces acting between laser beam and particle (reviewed by Mehta et al. (1999)).

Atomic force microscopes measure the deflection of a nanoscopic cantilever while it interacts with the sample. Figure 1.1 describes a typical experiment – a single protein is, at one end, picked up by the tip of the AFM cantilever while it remains fixed to a glass surface at the other end. The cantilever is then retracted from the surface and thus uncovers the forces that counteract the directed unfolding of the protein's domains (Fisher et al. 1999). These forces lead to the temporary deflection of the cantilever which is measured by laser interference. It remains difficult to control which part of the molecule remains attached to the surface or is picked up by the cantilever. For this reason, the technique is mostly applied to linear proteins that, naturally or by design, fold into chains of multiple repeating domains. Such an architecture is found in several "mechanical" proteins that have to withstand forces in their cellular context. The experiment then reveals the sequential unfolding of the domains that are situated between surface and cantilever (Rief et al. 1997; Kellermayer et al. 1997; Tskhovrebova et al. 1997; Rief et al. 1999). The atomic force microscope records a force-extension profile, that is the protein's force response at every point of elongation (see for example figure 2.3

in the following chapter). The sequential and independent unfolding of several domains leads to a typical saw-tooth pattern of force peaks alternating with the relaxation afforded by each unfolding event.

Single molecule techniques do not provide the same atomic detail as some of the more classic experiments described in the previous section. On the other hand, they may reveal stochastic fluctuations and substates that are averaged out in ensemble experiments. As I will show in chapter 2, atomic detail simulations can be crucial for the interpretation of such results. Protein dynamics appear to have much greater influence on the outcome of AFM experiments than has been previously assumed.

1.3 Theoretical methods for the study of protein dynamics

1.3.1 Molecular mechanics models

The properties of protein structures have been modeled at many different resolutions with different physical descriptions (Lazaridis and Karplus 2000). Insight into protein folding, for example, was gained from idealized lattice models which were, in early studies, not even meant to resemble any particular protein (Dill et al. 1995). Quantum mechanical models, at the other end of detail level, capture properties of individual electrons and can describe processes that involve chemical reactions (Gogonea et al. 2001; Gao and Truhlar 2002).

For the study of protein dynamics, atomic molecular mechanic models (reviewed by Wang et al. (2001)) constitute at the moment the perhaps best compromise between accuracy and computational cost. They are based on parameters that are independent of a particular molecule. Molecules are represented by spherical atoms of invariant radius and charge. A potential energy is then calculated depending on the exact position of each atom with respect to all others (Wang et al. 2001):

$$E_{total} = \sum_{bonds} K_r(r-r_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (1.4)$$

Equation 1.4 serves just as example. In detail, energy functions may differ between various simulation programs. However, they usually adhere to the general form shown above and differences remain mostly restricted to the parametrization, that is the exact combination of values for K, r_{eq}, θ_{eq}, and the other parameters. The total energy of the system is calculated as a sum of interactions between pairs of atoms. The first three terms in equation 1.4 determine the internal energy of the molecule. They penalize the deviation of chemical bonds from equilibrium length r_{eq}, equilibrium bond angle θ_{eq}, and equilibrium torsion angles, respectively (the latter described by energy barrier V_n, number of maxima n and offset γ). Parameters like r_{eq}, θ_{eq}, γ, n and V_n or the "spring constants" K_r, K_θ are usually derived from small molecule experiments or quantum mechanical calculations and depend on the types of atoms involved. Contrary to the previous three terms, the interactions considered in the last term do not follow chemical bonds but extend through space. They describe van der Waals (dispersion) and electrostatic attraction or repulsion and are computationally the most expensive to calculate.

The overall combination of energy function and parameterization is referred to as *force field*. Standard force fields make significant simplifications. The many body interaction of all atoms with all others is approximated by a two-body additive function. Furthermore, the constant charges centered on each atom reflect only roughly the nonisotropic charge distributions calculated from quantum mechanical models. The directionality of hydrogen bonds or effects of aromatic Π electrons are therefore not well reproduced. On the other hand, cooperative (many body) effects can to some extent be included implicitly into the parameterization of the energy function. After decades of development, different force fields are converging to similar representations (Wang et al. 2001). They are now routinely

applied to a variety of tasks with considerable success. Force fields are subject to ongoing development and recent formulations introduce polarization effects, that is they depart from the two body approximation (Cieplak et al. 2001; Kaminski et al. 2002).

Absolute energy values calculated from equations like 1.4 have no practical meaning but differences between them allow, for example, to compare different conformations of one molecule, to rank different orientations of a protein-ligand interaction or to (locally) optimize the geometry of a macromolecular structure. The probably most common application of such energy functions is the sampling of the conformational space that is available to a macromolecule. Molecular mechanical models are thus the primary computational tool for studying the dynamics of protein structure.

1.3.2 Molecular dynamics simulations

Equation describes 1.4 the energy E of a molecular system as a function of the atomic coordinates X. Driven by thermal fluctuations, the actual conformation evolves over time t across this potential energy surface and we would like to have a function $X(t)$ that describes this motion. The gradient of $E(X)$ with respect to the atom positions yields the forces acting on each atom:

$$F(X) = -\nabla E(X) \tag{1.5}$$

Newton's law translates these forces into accelerations and hence describes the dynamics of the molecular system:

$$F(X) = M\frac{d^2X(t)}{d^2t} \tag{1.6}$$

M is a diagonal matrix containing the mass of every atom. The solution of this differential equation yields the trajectory $X(t)$ of atomic positions over time. In practice, equation 1.6 becomes of course fairly complex and $X(t)$ has to be determined numerically. In order to perform this integration, simulation programs take a small step Δt in time, approximate the position x and velocity v of all atoms at

the end of this step from the current positions $x(t)$, forces $f(t)$ and velocities $v(t)$, and then recalculate these forces at the new positions. This task is for example implemented with the "Velocity Verlet" algorithm:

$$x(t+\Delta t) = x(t) + v(t)\Delta t + \frac{f(t)}{2m}(\Delta t)^2 \tag{1.7}$$

$$v(t+\Delta t) = v(t) + \frac{f(t)+f(t+\Delta t)}{2m}\Delta t \tag{1.8}$$

Equations 1.7 and 1.8 are given for a single atom with mass m. The second term of 1.8 averages the forces at time t and $t+\Delta t$ which increases the stability of the Verlet integrator. The whole procedure (1.7 and 1.8) is repeated until the simulation reaches the desired (or rather the computationally still affordable) time point. Accurate calculations require small time steps of 1 fs; Even a simulation that covers only 1 ns of protein fluctuation hence requires in the order of 10^6 iterations.

1.3.3 Different simulation regimes

In theory, the above equations fully describe the procedure of molecular dynamics simulations. However, in practice further technical details have to be considered to achieve a stable and efficient simulation. The van der Waals and electrostatic interactions in equation 1.4 need to be calculated for all pairs of atoms and the computational cost of this expensive cross term would thus increase exponentially with the size of the system. Nevertheless, nearly linear scaling is achieved if one ignores interactions between atoms that are separated by more than a certain cutoff distance. Inaccuracies due to this distance cutoff and the numeric integration itself may lead to the spurious loss or gain of energy in the course of a simulation whereas, in theory, the sum of kinetic and potential energy should remain conserved. This problem is often alleviated with a thermostat function which monitors the kinetic energy and adapts velocities so that the system's average temperature remains constant (Berendsen et al. 1984).

The representation of the solvent is another difficult issue. Real proteins are

surrounded by, and interact with, thousands of water molecules. Yet, for efficiency, one needs to limit the number of simulated particles as much as possible. Implicit solvent models offer one possible solution to this problem; Instead of explicitly adding water molecules to the simulated system, they emulate solvent effects by additional terms in the energy function (Roux and Simonson 1999). Alternatively, one may chose to surround the simulated protein by a layer or sphere of solvent molecules. The whole system is then subjected to stochastic friction and collision forces. This "Langevin dynamics" regime (Izaguirre et al. 2001) departs from the purely deterministic approach described so far but, at the same time, substitutes for a thermostat function. The third and arguably most realistic solvation strategy is based on a periodic boundary condition (PBC). The protein is placed in a box of explicitly modeled solvent molecules. The box is treated as the unit cell of a crystal. Particles crossing the boundary reappear on the opposite side of the box. Thanks to the periodic setup, Ewald summation can then be used for the calculation of long range electrostatic interactions between the particles in all cells (particle mesh Ewald summation, PME) (Essmann et al. 1995).

The work I present in the following two chapters makes use of all three solvation strategies. The unfolding of spectrin repeats was simulated with the generalized Born implicit solvent model (Bashford and Case 2000). It would have been impractical to maintain an equilibrated layer or box of explicit water while the protein's length was increasing from 4 to 24 nm. Under these circumstances, the implicit solvent model may well be more accurate than the less approximative explicit solvent methods; This will be further discussed in section 2.3.1. Moreover, Simmerling et al. (2002) recently employed the generalized Born approximation to simulate the folding of a small protein domain and their prediction agreed well with the subsequently solved experimental structure. Their work adds further support to the fidelity of implicit solvent methods and molecular dynamics simulations in general.

The conformational sampling of bound and free receptor and ligand proteins described in chapter 3 was performed in a layer of water molecules using Langevin dynamics. In this case, it was critical to find a realistic representation of the protein

surface as I was interested in the process of protein-protein interaction. A solvent layer with Langevin regime constitutes a compromise between the necessary detail level and the computational efficiency required for the study of 50 different (and often fairly large) proteins. This setup faired well for docking calculations and a mostly qualitative comparison of protein flexibility but proved insufficient for the reliable estimate of conformational entropies. I hence based the latter calculation on much longer simulations using the periodic boundary condition and PME treatment of electrostatics (see section 3.3.6).

1.3.4 Covariance analysis of simulations

Molecular dynamics simulations generate a wealth of data that needs to be analyzed. It may be instructive to simply watch the "movie" of atom movements but often one wants to extract a specific information. The fluctuating coordinates of several thousand atoms have hence to be translated into more manageable representations or quantities. In chapter 2 I examine the unfolding of spectrin repeats. I identified three important hinge regions within the molecule and the angles of these three hinges provided three intuitive dimensions for describing the trajectory.

However, often the coordinates of interest are much less obvious. Principle Component Analysis (PCA) and related techniques allow to decompose the data set into correlated motions and to reduce its dimensionality. The correlation of two atomic positions x_i and x_j (i.e. to which extend the two atoms move in concert) is measured by their *covariance*:

$$cov(x_i, x_j) = \left\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \right\rangle \tag{1.9}$$

A covariance of 0 implies two uncorrelated variables, i.e. the fluctuation of one atom is completely independent of the other's. If $cov(x_i, x_j) > 0$, atom i tends to move into the same direction as atom j or, if $cov(x_i, x_j) < 0$, their movement appears anti-correlated. From the simulated trajectory of atomic coordinates $X(t)$ we can derive a square *covariance matrix* C that contains the covariance of each

atom coordinate with respect to each other coordinate:

$$C_{ij} = cov(x_i, x_j) \tag{1.10}$$

The covariance matrix is subsequently decomposed into pairs of eigenvectors E_i and eigenvalues λ_i, for each of which holds $CE_i = \lambda_i E_i$. The eigenvector with the largest eigenvalue is called the (first) *principle component* and describes the direction of the largest collective motion within the system. The associated eigenvalue is a measure for how much of the molecule's fluctuation occurs along this important direction. More precisely, it quantifies the variance along this principle component. By concentrating on, for example, only the first two principle components, one can reduce the complexity of a molecular dynamics trajectory from $3N$ dimensions (i.e. the x, y, z coordinates of each atom) to the two artificial dimensions that retain the largest amount of information.

1.3.5 Entropy estimates

At first glance, a flexible molecule of N atoms possesses $3N$ degrees of freedom for moving through three dimensional space. However, in a protein, many atoms are quite rigidly attached to some others (directly or indirectly), which limits the degrees of freedom that are actually available to the molecule. The decomposition of the covariance matrix will hence reveal "vanishing" eigenvectors that are not independent and have an associated eigenvalue close to zero (i.e. zero variance). The number and magnitude of non-vanishing eigenvalues is hence related to the actual information content of the trajectory, which in turn is related to its entropy. Statistical mechanics defines this entropy S as:

$$S = -k_B \sum_n p_n \ln p_n \tag{1.11}$$

S hence depends on the number of microscopic states that are available to a system of atoms as well as on the probability p_n of each state. The more states exist and the more evenly they are occupied, the higher is the entropy of the system. k_B is the Boltzmann constant. The sum over a finite number of discrete states can be

replaced by the integral over a probability density function $p(x)$:

$$S = -k_B \int p(x)\ \ln p(x)\ dx \tag{1.12}$$

The integral can be solved analytically under the assumption that $p(x)$ is a 3N-dimensional Gauss distribution. The appropriate Gaussian fluctuations can be obtained by principle component analysis from a molecular dynamics trajectory. An exact derivation is beyond the scope of this introduction (and its author). As it turns out, the eigenvalue decomposition of covariance matrix C yields the solution to the problem (Karplus and Kushick 1981):

$$S = \frac{1}{2} k_B \ln |C| = \frac{1}{2} k_B \sum_{i}^{3N} \ln \lambda_i \tag{1.13}$$

Equation 1.13 should be extended by a correction factor (Schlitter 1993) that accounts for quantum mechanical effects.

A similar but not identical relation can also be derived directly from the quantum mechanical description of an harmonic oscillator. This "quasiharmonic analysis" examines the inverse of C and hence analyzes motions by their frequencies instead of amplitudes. Both methods utilize the (potentially rather anharmonic) motions of a molecular dynamics simulation but put them into a context that is, strictly speaking, only valid for harmonic vibrations. The simplification may be taken one step further and the same analysis can also be carried out without performing any simulation. Normal mode analysis constructs the Hessian matrix from the molecular mechanics force field and a single structure which is supposed to oscillate in a perfectly harmonic fashion (Case 1994). The covariance matrix C is the inverse of the Hessian matrix and their eigenvectors are identical. Quasiharmonic and normal mode analysis would be equivalent, if the protein's motions were indeed only harmonic.

In practice, quasiharmonic analysis should be more accurate but, unfortunately, it suffers from a sampling problem. The time scales accessible to computer simulation are still far from capturing the "complete" structure ensemble of a protein. A prolonged simulation will almost inevitably turn up additional con-

formation states that haven't been visited before and hence increase the calculated entropy (Gohlke and Case 2004). Nevertheless, as I will show in chapter 3.3.6, the entropy *difference* between free and bound state of a protein complex does indeed converge on sufficiently short time scales. Quasiharmonic analysis can therefore serve to estimate the conformational entropy lost or gained upon binding.

1.4 Function and dynamics – views in transition

1.4.1 Energy landscape of protein structure

Proteins fold into a unique ternary structure that is fully encoded in their amino acid sequence and arguably constitutes the thermodynamic minimum of the peptide chain within its solvent environment. This statement, Anfinsen's dogma (Epstein et al. 1963; Anfinsen 1973), could be considered one of the first important results and at the same time a foundation of early structural biology. It was based on the observation that some unfolded proteins could refold into their biologically active conformation outside the cell. Thus this conformation was (path)way independent. It could not rely on any specific process or information in the cell and must constitute a thermodynamic minimum. The latter conclusion was countered by Levinthal's paradox (Levinthal 1968): A unique protein structure could not be reached by a simple random search. Thus protein folding must follow a sequence of events, one or several pathways, and the native conformation could still constitute a local minimum, a kinetically trapped state. This dilemma captured the interest of many structural biologists who tried to delineate folding pathways to unique protein structures.

Another line of research focused on the structural basis of protein function, especially the mechanism of enzyme specificity, catalysis and regulation. This study of proteins "in action" increasingly abandoned the notion of a unique protein structure. Koshland (1958) postulated protein flexibility and introduced the idea of an induced fit to explain the recognition between enzymes and their substrates. Monod et al. (1965) explained the cooperative regulation of multimeric enzymes

by two distinct protein conformational states with different affinities for each other as well as for the substrate. Kinetic experiments on the binding of substrate to enzyme (Kirschner et al. 1966) or carbon monoxide to myoglobin (Austin et al. 1975) supported the assumption of distinct conformational substates. The study of atomic fluctuations with crystallographic B-factors and Mössbauer spectroscopy added further diversity to the picture and revealed different motional regimes. Beyond a transition temperature of about 200 K, mostly vibrational fluctuations are superseded by larger scale anharmonic movements and the latter are required for protein function (Parak et al. 1982). Proteins appear to explore a continuum of conformations rather than a finite set of states.

This modern view is summarized by the concept of a 3N-dimensional energy landscape (Frauenfelder et al. 1991). A system of N protein (and solvating) atoms moves through a hyperspace of free energy wells and barriers. The landscape is rough at every level of resolution – there are wells within wells within wells (Ansari et al. 1985). At the same time, it is sufficiently biased toward the native state to ensure protein folding (Dill and Chan 1997). Wells and barriers are organized in hierarchical fashion. The global folding funnel may lead to several functional substates, which are separated by high free energy barriers but are themselves exploring a collection of subconformations (Fenimore et al. 2004). The concept of an energy landscape thus reunited the research on protein folding with the the study of protein function and dynamics. It resolved Levinthal's paradox without violating Anfinsen's dogma.

Rather than as single structure, proteins should hence be thought of as ensemble (or distribution) of structures. The origin and implications of protein motions are a matter of ongoing research (Fenimore et al. 2002; Fenimore et al. 2004). To which extent is protein function influenced by molecular fluctuations? For some proteins, the connection between function and dynamics is evident – for example, molecular motors have the very purpose of implementing large scale movements (Howard 1997) and their mechanism of action involves molecular fluctuations on many scales (Astumian and Bier 1994; Schief and Howard 2001). However, in most other cases, such a connection is less obvious and not well understood.

1.4.2 Directed unfolding of proteins

Experiments on the forced unfolding of single molecules were early on accompanied by molecular dynamics simulations (Lu et al. 1998; Krammer et al. 1999; Lu and Schulten 1999; Marszalek et al. 1999; Paci and Karplus 1999; Paci and Karplus 2000; Lu and Schulten 2000; Best et al. 2001). One could hence expect protein flexibility to be well considered in our understanding of force-induced unfolding. Yet, as I will discuss in section 2.1.2, the experiments and consequently also simulations had initially concentrated on one family of highly resistant protein domains that unfold in a singular well defined event. Such regular behavior was subsequently also expected from other proteins. For this reason, results of atomic force microscopy experiments were generally interpreted with models that assumed single-step unfolding along a unique unfolding pathway. Studies were mainly concerned with the connection between (static) molecular architecture and unfolding route. Atomic fluctuations seemed less relevant.

This static framework was put into question by atomic force microscopy experiments on a different mechanical protein domain. The (controversial) force response of a spectrin repeat suggested different types of unfolding events and contradicted single pathway, all-or-none unfolding. Chapter 2 describes these experiments and corresponding simulations. An ensemble view on protein structure turns out to be instrumental for the understanding of spectrin behavior and function.

1.4.3 Protein-protein binding

Binding processes were among the first to reveal the importance of molecular fluctuations for protein function. Fast kinetic studies of glyceraldehyde-3-phosphate dehydrogenase activity, for example, showed that substrate binding is influenced by a (comparatively) slow exchange between different conformational states of the protein (Kirschner et al. 1966). Myoglobin provides another example for the obvious necessity of molecular movement. Its binding site for oxygen and carbon monoxide lies buried within the molecule and lacks any opening to the

solvent. Austin et al. (1975) analyzed the temperature dependence of CO binding and attributed one (out of four) free energy barriers to the equilibrium between many conformational states. Later, Ansari et al. (1985) identified a hierarchy of relaxation motions during the dissociation of CO from myoglobin. Similar observations were also made for the interaction between proteins and antibodies (e.g. Lancet and Pecht (1976), Foote and Milstein (1994)). Molecular dynamics simulations were in some cases used to examine the free energy profile for the force-induced dissociation of small molecules from proteins (Izrailev et al. 1997; Balsera et al. 1997) and Heymann and Grubmüller (2001) provided an explicit estimate for the conformational entropy along such an unbinding trajectory.

Nevertheless, the interplay between structural dynamics and protein-protein recognition is still poorly understood. In general, it remains unclear to which extent protein flexibility affects the kinetics and thermodynamics of binding. Computational descriptions of protein-protein interaction treat the binding partners as completely or mostly rigid bodies. Common kinetic models for the binding process are based on the same assumption. Furthermore, there are conflicting results and predictions as to the influence of conformational diversity (and thus entropy) on the thermodynamic stability of protein-protein complexes. The effective treatment of flexibility is currently also the highest obstacle to the reliable prediction of protein complexes from their single components. Chapter 3 (in particular sections 3.1.2 to 3.1.4) provides an in-depth discussions of these issues.

1.5 Conclusion

Proteins are not static but constantly on the move. The complex dynamics of these molecules is often the missing link between structure and function. Its study requires the integration of various experimental and theoretical data. The following two chapters describe how a detailed view on protein dynamics can, on the one hand, provide insight into individual protein function and, on the other hand, improve our understanding of a fundamental biological process.

Chapter 2

Forced unfolding of spectrin repeats

2.1 Introduction

2.1.1 The spectrin repeat – a domain under stress

Twenty five thousand billion red blood cells circulate through several hundred kilometers of human capillary network. With their diameter of 7 μm they have to squeeze through capillaries that are only 3 to 5 μm wide. The ability to accommodate large reversible deformations is hence a crucial property of erythrocytes. This remarkable elasticity is conferred by the membrane skeleton – a two-dimensional protein mesh work that lines the inner face of the plasma membrane (Discher and Carl 2001) (figure 2.1A). Most cell types rely on this membrane skeleton to stabilize and spatially organize membranes while a three-dimensional cytoskeleton secures their shape integrity (De Matteis and Morrow 1998). However, the elastic erythrocytes lack a rigid actin-based cytoskeleton and their membrane skeleton thus also maintains the cell's overall shape (Elgsaeter et al. 1986; Hansen et al. 1996). The same holds for the outer hair cells of the ear whose function also crucially depends on shape elasticity(Raphael et al. 2000).

The predominant component of the membrane skeleton is spectrin, an elongated multi-domain protein composed of two α and two β subunits (figure 2.1B). Experimental evidence and theoretical calculations indicate that these $\alpha_2\beta_2$ te-

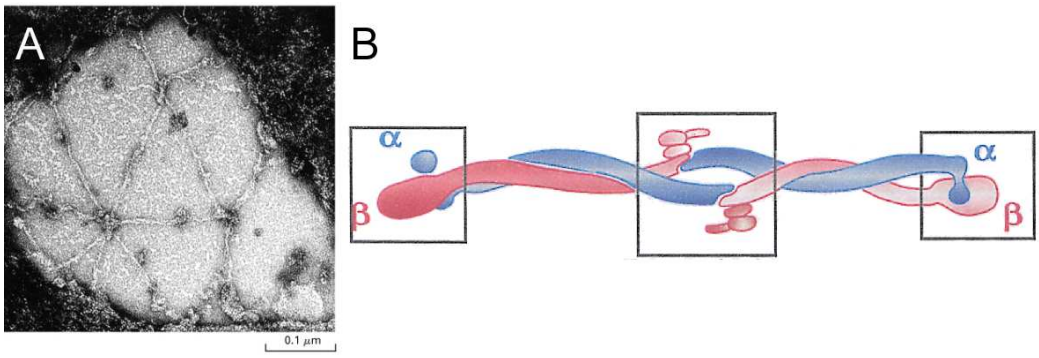

Figure 2.1: Spectrin tetramers are the "connectors" of the two-dimensional membrane skeleton underneath the plasma membrane. A) Electron micrograph of the erythrocyte membrane skeleton (courtesy of Daniel Branton). B) Schematic view of the spectrin tetramer. Boxes mark the dimerization (outer rectangles) and tetramerization (inner rectangle) sites, respectively. [reproduced with kind permission from Byers and Branton (1985) (A) and Pascual et al. (1997) (B)]

tramers are the elastic part of the network (Byers and Branton 1985; McGough 1999; Lee and Discher 2001). Both subunits are largely made up from a series of spectrin repeats, each of which typically contains 106 amino acids. Each spectrin repeat (figure 2.2) is built from three anti-parallel α-helices, which are separated by loops and fold into a left-handed coiled-coil (Pascual et al. 1997; Djinovic-Carugo et al. 1999; Grum et al. 1999).

The spectrin repeat is one of the most abundant domains in the human genome. About 500 spectrin repeats can be found, most commonly in actin filament associated proteins such as spectrin itself, dystrophin, utrophin, and alpha-actinins (Pascual et al. 1997). The domain usually occurs in stretches of 4-40 repeats. Interestingly, many of the proteins containing spectrin repeats are found in locations that are regularly subjected to mechanical stress, not only in the membrane skeleton but, for instance, also the muscle Z band and muscle-basement membrane contacts (Pascual et al. 1997).

Part of spectrin's elasticity could be explained by the network it forms and by modest shifts in its connectivity (Lee and Discher 2001). Changes in the super coiling of spectrin's α and β subunits could provide elasticity on the level of the single tetramer (McGough 1999). Changes in the interactions between par-

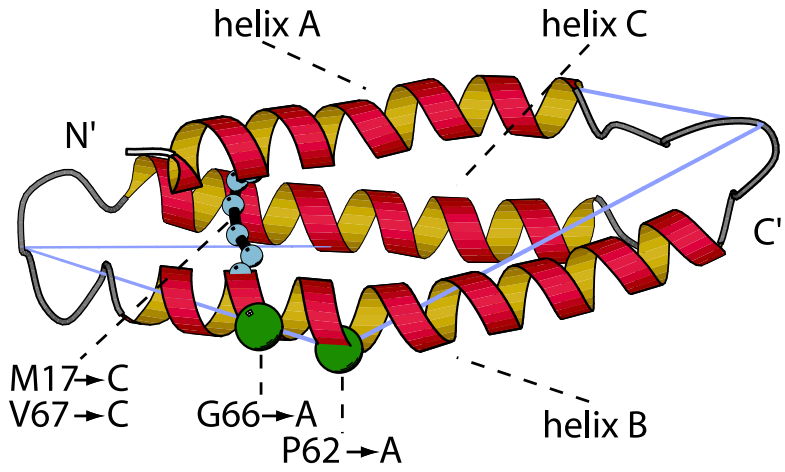

Figure 2.2: Structure of the spectrin repeat R16. C_α atoms of the two residues substituted in the AA mutant are highlighted in green. The disulfide bond introduced into the CC mutant is marked in blue. Blue lines describe the three angles that will be used to represent hinge movements in figures 2.5 and 2.10.

ticular domains are very likely contributing another level of elasticity and there is evidence of stabilizing interaction between neighboring spectrin repeats (MacDonald and Pozharski 2001). Less clear is whether and how the single repeat may contribute to the mechanical properties of proteins containing it.

2.1.2 Forced unfolding of spectrin repeats and other domains

Already before the onset of our collaboration, the lab of Heinrich Hörber had studied the mechanical properties of the spectrin repeat using atomic force microscopy (AFM, briefly introduced in section 1.2.3). They had engineered constructs consisting of four identical copies of the 16^{th} repeat of α-spectrin and stretched single molecules of this type between a surface and the tip of an AFM cantilever (Lenne et al. 2000). Some of their results were at odds with the behavior commonly expected for the forced unfolding of "mechanical" protein domains.

Since the original work by Rief et al. (1997) on titin it had become widely accepted that single domains of such linear proteins will unfold in an all-or-none fashion when subjected to the directed force of the AFM. However, most of these

studies (Carrion-Vazquez et al. 1999; Li et al. 2000; Oberhauser et al. 2001; Carrion-Vazquez et al. 1999) had focused on only two domains that have to sustain high forces in their physiological environment. Both immunoglobulin (Ig) type I domains and the fibronectin type III (FnIII) domain are all-β proteins and have a similar (Ig-like) topology. Both domains are found in the giant muscle protein titin. Another well studied example, the extracellular matrix protein tenascin (Oberhauser et al. 1998), also consists mainly of FnIII domains. Molecular dynamics (MD) simulations performed on these β-sandwich structures of Ig-like domains reinforced the notion that globular protein domains support only little deformation under external force. A catastrophic event then transforms the proteins rapidly into an unordered polypeptide peptide chain with little or no higher structure left (Krammer et al. 1999; Lu et al. 1998; Lu and Schulten 1999; Lu and Schulten 2000). This view was revised by Marszalek et al. (1999) who suggested that IgI domains may display a transient kinetic intermediate during constant speed forced unfolding, based on experimental data supported by MD simulations. However, since this intermediate occurred immediately before the catastrophic event after only a few Å of deformation, it did not really pull the notion of all-or-none unfolding into question (Lu and Schulten 2000).

The observations of Lenne et al. (2000) on spectrin repeats stood out from this previous work on Ig-like domains in three respects: Firstly, the forces needed for the disruption of single spectrin repeats were markedly lower than the forces measured for IgI and FnIII domains. Typical forces ranged between 50 and 80 pN compared to 200 pN in the case of, for example, IgI. Secondly, and perhaps most surprising, Lenne et al. reported two distinct populations of unfolding events. Apart from the expected extension by 32 nm which corresponds to the complete unfolding of a single spectrin repeat, they found a number of events in which the molecule gained only about 15 nm length. They suggested that these events could represent the break up of a putative unfolding intermediate. Thirdly, the unfolding lengths recorded by Lenne et al. varied broadly around the 32 and 15 nm average values and also the rupture forces were subject to considerable variation. Ig-like domains, by contrast, showed nearly constant extensions and rupture forces in previous studies (Carrion-Vazquez et al. 1999; Oberhauser et al.

1998; Rief et al. 1997). Indeed, such behavior was considered standard to the extend that the established AFM protocol prescribed the filtering of experimental force-extension profiles for regular peak spacing (Rief et al. 1999).

The results of Lenne et al. thus contradicted assumptions commonly made by AFM experimentalists. However, the universal validity of all-or-none unfolding was also questioned by one later experimental and two concurrent theoretical studies. Best et al. (2001) demonstrated in a study on forced unfolding of barnase that such regular behavior could not necessarily be expected from "non-mechanical" proteins. Moreover, simulations by Paci and Karplus (1999, 2000) showed that β-sandwich as well as a α-helical structures could theoretically create stable forced unfolding intermediates at much larger extensions.

2.1.3 Our approach

The study of Lenne et al. attracted critics because it arrived at unexpected results using unconventional methods. An unfolding event at 15 nm extension implied an intermediate that was either 15 nm longer than the intact – or 15 nm shorter than the completely stretched domain. Both possibilities reduce to a position in the middle of complete unfolding (which yields 32 nm). It was difficult to imagine, how the simple topology of the spectrin repeat should support a metastable state that was roughly 6 times longer than the native fold. Furthermore, Lenne et al. did not filter their force curves for regular peak spacing (which would have ruled out the detection of unfolding intermediates) but tried to exclude experimental artifacts with a different protocol.

In this situation, the groups of Matti Saraste, Heinrich Hörber and Michael Nilges entered a collaboration to study the problem by a combination of experimental and computational methods. Jari Ylänne and Kristina Herbert expressed and purified two constructs containing mutated spectrin repeats. Pierre-François Lenne, with the assistance of Stephan Altmann, subjected wild-type and mutant spectrin repeats to AFM experiments. I examined the forced unfolding of wild-type and mutant spectrin repeats with molecular dynamics simulations and tried to relate experimental and theoretical results.

The comparison with simulations could indeed explain the main aspects of the experimental unfolding data. The lower rupture forces were traced to the α-helical composition of the spectrin repeat and the fact that unfolding breaks atomic interactions in a step by step manner. The broad variation of experimental unfolding lengths was mirrored by a surprisingly variable rupture point of the simulated domain. Across different simulations, the native fold unraveled at very different extensions that were often several nm apart. The simulations also showed potential unfolding intermediates of appropriate length. These intermediates arose from non-native topologies and depended on a pronounced kink of helix B in the center of the molecule. My collaborators engineered a variant of the spectrin repeat with a strengthened helix B and this mutation indeed prevented the occurrence of unfolding intermediates.

Both a "fuzzy" breaking point and metastable "blocks" along some unfolding pathways can sum up to a smooth and elastic response to external force – if we consider that spectrin repeats always occur in a series of many replicas. Each domain is moving through conformational space and none of the replicas have exactly the same structure at any given moment. The same force can thus provoke different responses from perfectly identical domains. This ensemble effect appears to be the key to understanding both the experimental results and the biological function of the spectrin repeat.

The following sections first describe the experiments and simulations on wild-type spectrin repeats. I then discuss and explain the peculiarities of the experimental results using the molecular picture gained from the simulations. Support for this interpretation is subsequently drawn from the examination of the two modified spectrin repeats. Finally, I compare our results with previous studies on the matter. The chapter concludes with a summary putting the outcome of our work into the biological context of spectrin function.

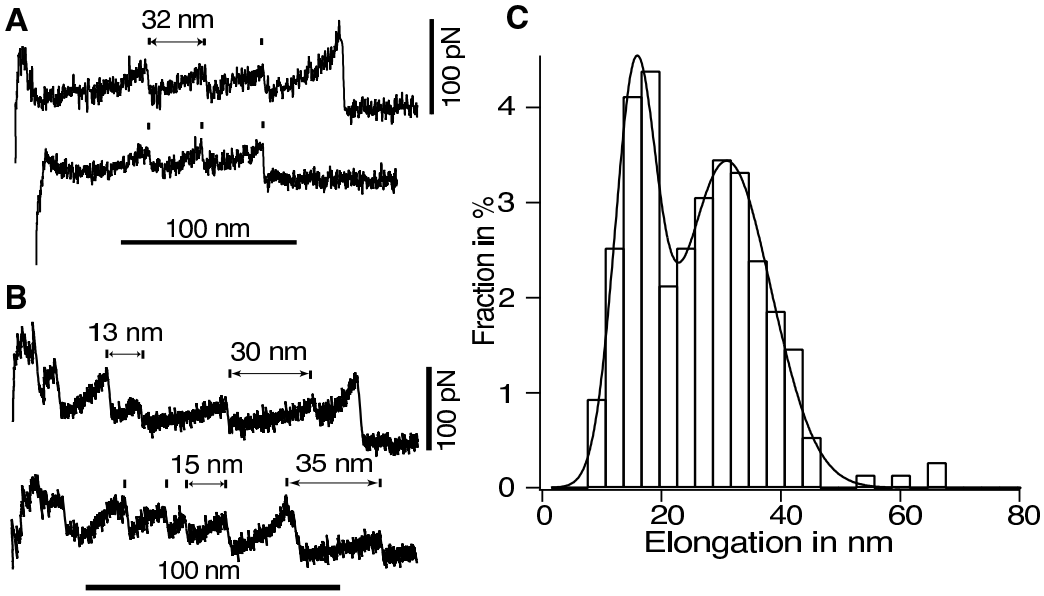

Figure 2.3: Forced unfolding experiments on the wild-type spectrin repeat. A peptide chain made up of 4 identical repeats was stretched and the restoring force measured. A) A fraction of curves display regular unfolding lengths around 32 nm. B) The majority of curves show varying unfolding lengths with both intermediate and full-length peak-to-peak distances. C) Distribution of elongation lengths from 250 unfolding experiments.

2.2 Forced unfolding of wild-type spectrin repeats

2.2.1 Atomic force microscopy experiments

In AFM experiments, a single protein is stretched between a surface and the tip of an microscopic cantilever and the elongations as well as the forces necessary for the deformation are measured in real time down to the millisecond time-scale. A peak in the AFM force-extension profile marks the break-up of a stable conformation of the peptide chain. In a protein construct with multiple identical domains, the repeated saw-tooth pattern in the AFM force-extension curve is an indication of repeated unfolding events, specific for the domains contained in the construct. The spacing between peaks corresponds to the gain in length due to the unfolding of a part of the protein chain.

As before, P.-F. Lenne and S.A. Altmann performed their experiments with

a protein construct containing 4 identical copies of the 16^{th} repeat from chicken α-spectrin. They used an elongation speed of 0.3 nm ms^{-1}. To rule out any interaction between the domains, the repeats were each separated by an additional (presumably α-helical) linker of 17 amino acids. The new measurements confirmed the original findings of Lenne et al. (2000). The distribution of rupture forces was quite broad and peaked between 50-80 pN (data not shown). Examples of force curves are given in figure 2.3. As indicated, my collaborators observed consecutive unfolding events with various peak-to-peak distances. They did find regular curves that represent the sequential unfolding of complete domains each yielding 32 nm (figure 2.3A). However, the majority of experiments showed significant variation of peak-to-peak distances. The histogram in figure 2.3C highlights this variation. Two representative examples are given in figure 2.3B. Five unfolding events were counted in the second curve, i.e. one more than the number of domains in the construct, while the total length of unfolding is less than the maximum stretched length of the construct. As Lenne et al. (2000) already reported and as figure 2.3B illustrates, shorter peak-to-peak distances were often detected in the beginning of the experiments.

The force-extension profiles revealed a complex unfolding behavior of the tetrameric spectrin construct. The wide and bi-modal distribution of peak-to-peak distances could not be explained by independent all-or-none unfolding of complete spectrin domains. On the contrary, single force curves already exhibited variations from the average distance that appeared beyond experimental error and the histogram of peak-to-peak distances (figure 2.3C) indicated a significant fraction of partial unfolding events.

2.2.2 Steered molecular dynamics simulations

I simulated the mechanical unfolding of the wild-type spectrin repeat with a constant extension rate of 0.2 Å ps^{-1} (2×10^7 nm ms^{-1}) by molecular dynamics. The molecule was extended to a distance r_{NC} between the N and C termini of about 24 nm which corresponds to an elongation of the native structure by 20 nm. This value is smaller than the molecule's maximal extension as I was mainly interested

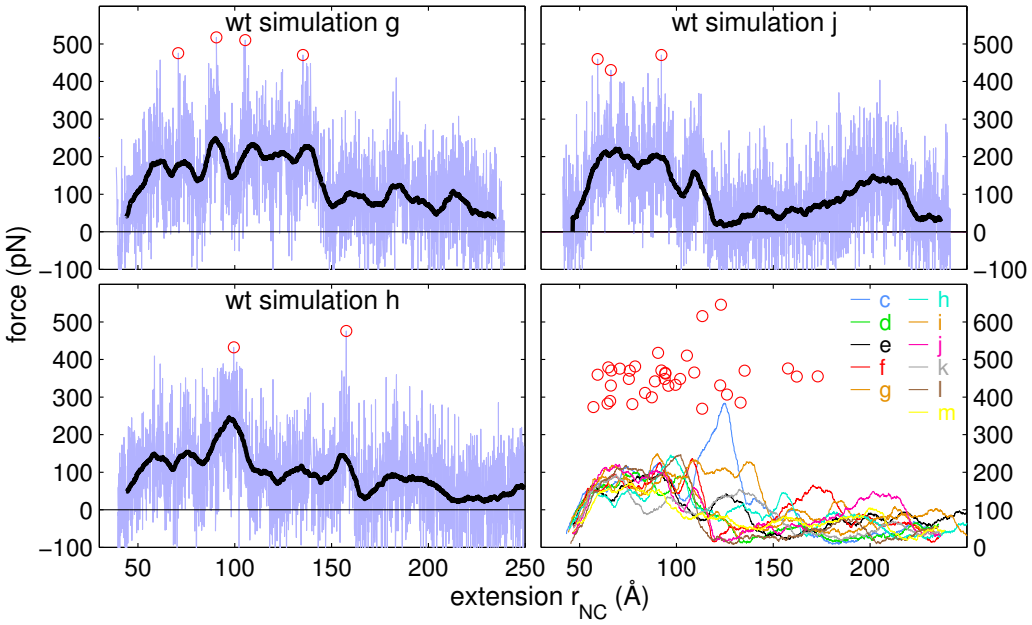

Figure 2.4: Force-extension profiles from unfolding simulations with the wild-type spectrin repeat. Three profiles are shown in detail. The complete data recorded at 0.5 ps resolution are drawn in blue. Black is a running average with 50 ps window size. The running averages of all simulations are presented in the lower right plot (note the shifted force scale). Force peaks within 10% of a trajectory's maximum force are marked with circles and are included for all simulations in the lower right plot. They illustrate the variance of extensions sustained by the repeat.

in the rupture of the tertiary structure. I calculated 11 trajectories of 1 ns each. Unfolding was enforced by a time-dependent harmonic distance restraint between N- and C terminus. The molecule's resistance toward unfolding is illustrated in the force extension profiles of figure 2.4. In analogy to the displacement of the experimental cantilever, restoring forces were calculated from the deviation between r_{NC} and the target value to which it was restrained.

The maximum force recorded in each trajectory showed a broad variation from 400 to 645 pN with an average of 475 pN ($\pm$65 pN). In contrast to other systems studied by AFM and MD (Paci and Karplus 1999; Lu and Schulten 2000; Lu and Schulten 1999), the extension at which this peak force occurred was not well defined. It varied between 6.5 and 16 nm r_{NC}. Moreover, several peaks of sim-

ilar force were often observed at different extensions. In figure 2.4, I arbitrarily marked all forces within 10% of each trajectory's maximum. The varied positioning of such peaks at extensions up to 11 nm apart has important implications for the variation of unfolding lengths in AFM experiments.

Unfolding was typically initiated by a gradual stretching of the two outer helices A and C, with varying restoring forces. Hence, the native triple-helical coiled coil topology sustained considerable elongations. This phase culminated in the disruption of the native fold, marked by a drop in the force extension profile. Several unfolding trajectories proceeded without further pronounced resistance (simulations d, i, l, and m; see figure 2.4). However, six simulations exhibited distinct force peaks after the initial unfolding event and, in two cases (e.g., simulation h) such peaks constituted the maximum force of the whole simulation. The late force peaks were caused by compact but non-native folds that provided interim mechanical stability at various extensions.

In order to describe the origin of these potential intermediates I focus on shifts in the molecule's topology rather than analyzing the complex atomic detail of my simulations. The native spectrin repeat consists of three secondary structure elements – helices A, B and C – which are connected by two loops – AB and BC (figure 2.2). The two loops are the obvious hinge points in the repeat's topology that need to open up to completely unfold the molecule. Before that, however, the simulations showed a limited melting of helix B near Pro62 in the center of the molecule. Increasing a naturally occurring bend in this position, the central helix was effectively divided into two. Only this additional hinge allowed the repeat to evade complete unfolding beyond lengths of about 13 nm r_{NC}.

In figure 2.5 the status of each of these three hinges is described by a separate axis. Each conformation of a given trajectory corresponds to a point in this three-dimensional graph. Transitions along axis AB indicate opening of loop AB, transitions along axis BC show the disruption of loop BC and the third axis B describes the closing and opening of the additional hinge in helix B. We can thus follow the topology of the repeat in the course of each trajectory. In addition, I color-coded the smoothed force profile from figure 2.4 onto each trajectory trace

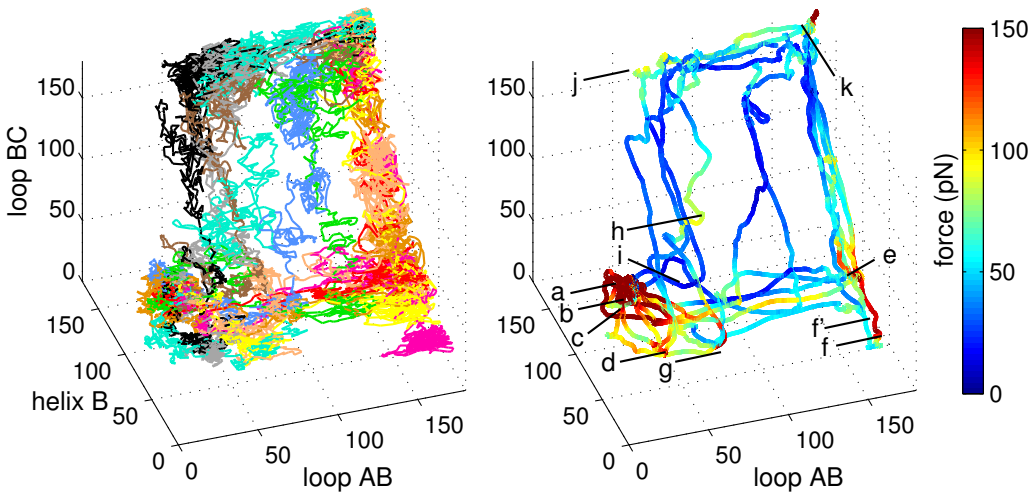

Figure 2.5: Topology and force resistance of the unfolding wild-type spectrin repeat. Trajectories are described by three hinge movements (see figure 2.2). The complete recording is shown on the left where trajectories are colored as in figure 2.4. For the right plot angles and forces were averaged over a 50 ps window and forces then color-coded onto the smoothed topology traces. Conformations exhibiting high restoring forces are prominent around the native structure (a) and the completely stretched end point (k). Moreover, I observed two force resistant non-native topologies (d, g and e, f). Structure snapshots for each topology region are marked with letters and given in figure 2.6.

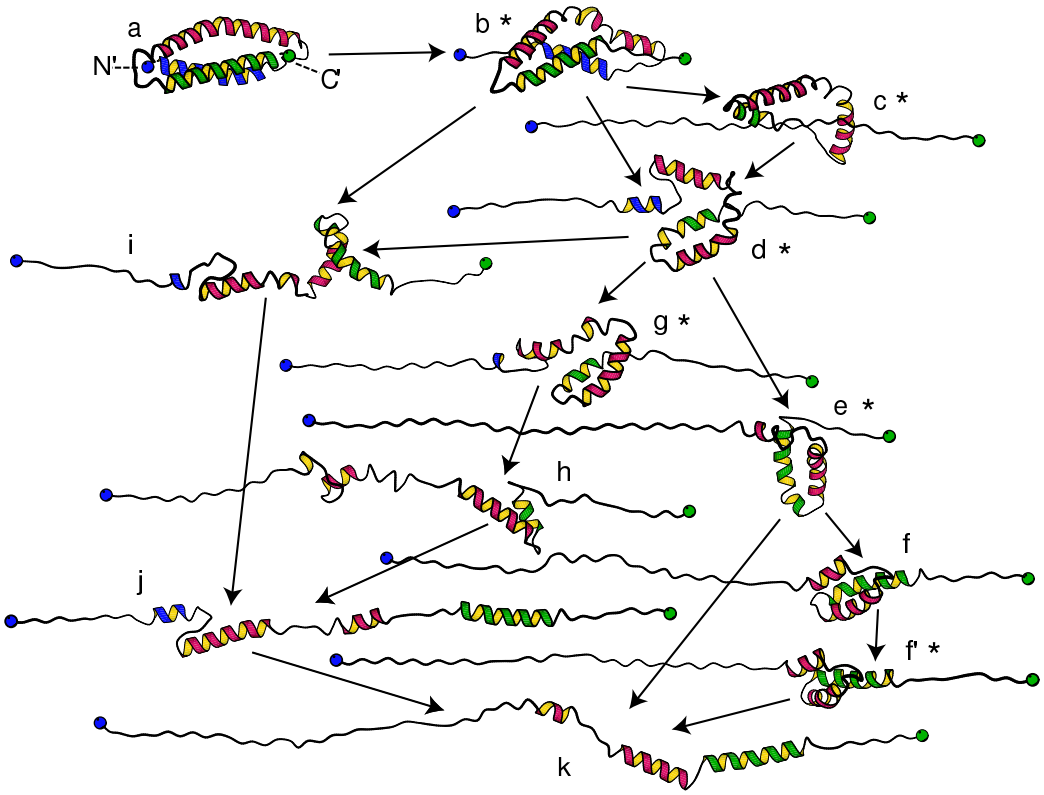

Figure 2.6: Characteristic topologies of the unfolding spectrin repeat. Snapshots were chosen to describe prominent regions of topology from figure 2.5. Structures marked with * are directly corresponding to force peaks. Simulations started from the NMR structure a. Snapshots b and c were taken from simulation g; f, f', k from simulation j; g and h from simulation h. The detailed force profiles of these three simulations are given in figure 2.4. The structures have the following lengths (r_{NC} in Å): a 42, b 71, c 135, d 126, e 173, f 191, f' 205, g 158, h 175, i 143, j 192, k 241.

(blue for lowest forces; red for highest forces). From the resulting plot one can easily pick out (1) topologies which exhibited pronounced force resistance and (2) how often these topologies were visited in the 11 simulations.

Two main routes emerge from a variety of paths through the simplified topology landscape. Along these routes I identified 4 regions of distinct topologies. Example structures for each of these topologies are marked with letters and shown in figure 2.6. The first region (illustrated by snapshots a, b, and c) portraits the native fold of the spectrin repeat. This native topology resisted unfolding over a wide range of elongations. A prominent second region of force resistant topologies is depicted by snapshots d and g. It features the sharp kink in helix B and appears hence shifted along axis B in figure 2.5. Due to this kink, compact structures formed from the remaining part of helix C, the two halves of helix B, and a distorted but still persisting loop AB. In terms of extension this region partly overlaps with the longest observed native folds. Snapshots c and d provide an example of this extension range where high unfolding resistance could stem from a still native-like topology in one simulation whereas similar resistance was evoked by the non-native arrangement in 5 other simulation runs.

The two remaining regions of prominent topology belong to the two dominant unfolding pathways. In three trajectories loop BC was disrupted first and the kink in helix B was straightened. Loop AB persisted over a considerable range of extensions (snapshot j) but this conformation displayed only moderate resistance to further unfolding. In the other, most frequent pathway, loop AB opened up first. The kink in helix B allowed for pronounced force peaks from a helix-loop-helix element (snapshot e) consisting of the remaining parts of helix B and C locked together by the native BC loop. The longest intermediate observed originated from this structure and had its maximum unfolding resistance at 21 nm r_{NC}. In this particular case, the helix-loop-helix element was tightly folded back onto the N terminal half of helix B and the resulting (non-native) fold resembled spectrin's native triple helical coiled coil (snapshot f'). The detailed force profile of this simulation (j) is given in figure 2.4.

Three trajectories followed neither of the two pathways strictly but left the

initially chosen main route by a premature opening of loop BC. I prolonged two simulations to complete extension of the molecule, marked by a steep linear increase of restoring forces (not shown). In experimental reality, unraveling domains can only be extended up to the disruption of the next folded structure, that is the weakest element of the chain. In my simulations I assumed this to be the case if a running average of restoring forces (as shown in figure 2.4) surpassed a threshold of 300 pN. The resulting maximal distances between N and C terminus were 35.8 and 35.9 nm with the actual numbers being rather insensitive to the choice of definition or force threshold. Due to the initial distance between N and C termini (4 nm) this would correspond to a length gain of 32 nm in AFM experiments.

2.3 From experiment to simulation and back

2.3.1 Translating between simulation and experiment

There are certain difficulties in comparing the simulations and AFM experiments. First, because of the variability encountered, 11 simulations for the wild type repeat seem not sufficient to obtain statistics comparable to the experiment. Second, I only simulated a single domain and in a polymer of 4 spectrin repeats used in the AFM experiments the status of the completely or partially folded but prestretched domains is undefined at the time of single-domain rupture. Furthermore, it is unknown to which extent the partially distorted domains refold when the force drops. Third, similar to other studies (Krammer et al. 1999; Lu et al. 1998; Lu and Schulten 1999; Lu and Schulten 2000; Best et al. 2001; Craig et al. 2001), elongation speeds of experiment and simulation differ by 8 orders of magnitude (3×10^{-7} versus 20 m s^{-1}). Despite this discouraging gap between experimental and computationally accessible time scale, MD simulations have already been instrumental for the interpretation of unfolding experiments and, in various instances, provided results consistent with AFM data (Li et al. 2000; Marszalek et al. 1999). Simulations, in general, reproduce an unfolding scenario regardless

of pulling speed or force and formulation of the restraint (Lu and Schulten 1999; Paci and Karplus 1999; Izrailev et al. 1997). Even quantitative statements about the height of energy barriers have been attempted (Izrailev et al. 1997) which took into account that simulations operate much further away from equilibrium than AFM experiments (Balsera et al. 1997).

Lu and Schulten (2000) criticized the application of implicit solvent models (such as the generalized Born continuous solvent model I employed) to unfolding simulations since they cannot describe the detailed mechanism of (concurrent) hydrogen bond breaking. On the other hand, Paci and Karplus (1999) have argued that an implicit solvent model alleviates artifacts of high unfolding speeds in the simulation since it mimics equilibrated water at each step of the simulation. Apart from practical considerations (size of the water bath necessary to simulate the complete unfolding of a spectrin repeat, necessary CPU time) the approximation may well be an advantage since a slow step in the unfolding (equilibration of water around the unfolding protein) is not simulated. The generalized Born model has been demonstrated to be a good approximation to explicit solvent simulations of proteins; see, for example, Cornell et al. (2001).

In summary, despite of the difficulties in direct comparison between AFM experiments and MD simulations, lower unfolding forces and the broad distribution of force peak positions could be seen in both. Moreover, MD simulations suggested unfolding pathways and provided a way to experimentally test these pathways. This is further discussed below.

2.3.2 Low unfolding forces

As already mentioned, the work on mechanical properties of proteins had so far largely concentrated on the muscular protein titin (reviewed by Linke and Granzier (1998)). In titin, IgI and FnIII domains appear to unfold only at high forces above the physiological range, as they may occur, for example, in over-stretched muscle filaments (Linke and Granzier 1998). Although the matter is still under debate (Minajeva et al. 2001), individual Ig-like domains seem to function mostly as “emergency reserve” rather than as elastic elements. The molecular

architecture of IgI and FnIII domains reflects this function. A "seal" of several hydrogen bonds (6 in the best studied case) between two anti-parallel β-sheets protects the native fold from extension by force (Lu and Schulten 2000). Unfolding requires the simultaneous rupture of these hydrogen bonds and is hence only achieved by high forces.

By contrast, spectrin is built exclusively from α-helices arranged parallel to the force axis. Backbone hydrogen bonds are hence connecting neighboring residues only and can be disrupted one after the other. This step-wise unfolding (and in some simulations also partial refolding) of the outer helices buffers the impact of an external force on the spectrin fold. Peak forces of the simulations are associated with the complete or partial disruption or re-arrangement of the native helix arrangement. The native fold is stabilized by 3 clusters of hydrophobic side chains, each of which connects the 3 helices perpendicular to the force axis. Interactions of this type are, compared to the hydrogen bond patch of IgI, weaker but also more tolerant to re-arrangements. The peak forces of spectrin simulations are thus lower than the forces from similar simulations on Ig-like domains (Marszalek et al. 1999; Paci and Karplus 1999; Craig et al. 2001) – in line with the experimental observations. However, forces vary with unfolding speed and simulation conditions. For exact statements one would need to compare the unfolding of the different domains with identical simulation setup. Paci and Karplus (2000) have performed such simulations and showed that spectrin repeats unfold at forces below IgI and FnIII domains but above the restoring forces of an α-helical domain without mechanical function.

2.3.3 The variation of unfolding lengths

Peaks or rather the sudden drop of force in AFM experiments mark the break-up of a single domain (Carrion-Vazquez et al. 2000). In a chain consisting of identical repeats, all of them should likewise be on the verge of breaking at the time of an unfolding event. Hence, the distance between two force peaks in the experiment corresponds to the gain of length due to unraveling of the folded content of a *pre-stretched* domain (as pointed out by Lu and Schulten (1999)).

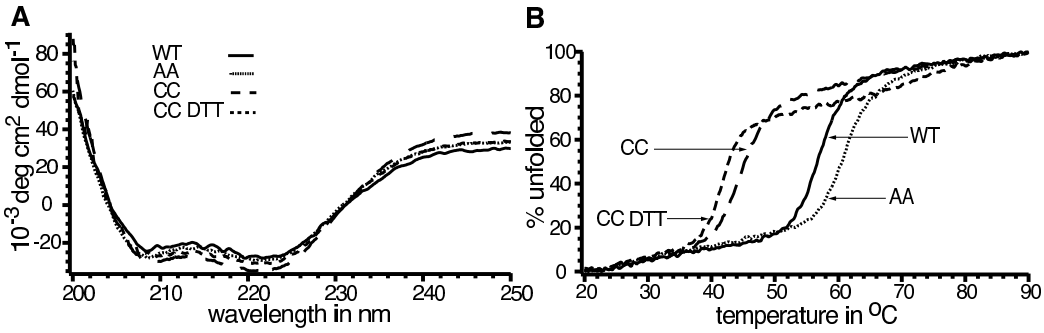

Figure 2.7: Stability of wild type and mutant spectrin repeats. A) CD spectra of spectrin repeats at room temperature: wild type (WT), double alanine (AA) and the double cysteine mutants without (CC) and with 10 mM DTT (CC DTT). B) Percentage of unfolded values during a temperature ramp monitored at 222 nm for these repeats. The thermal denaturation was not reversible.

A common feature of all AFM experiments was the broad distribution of these peak-to-peak distances. Such variation might be, in part, attributed to the low force needed for a single unfolding event (50-80 pN) which required my collaborators to work close to the sensitivity limit of the AFM instrument. A modification of the spectrin repeat allowed them to test the actual sensitivity of the AFM experiments: Jari Ylänne and Kristina Herbert replaced two amino acids of the native domain by cysteine residues in order to connect helices A and C with an artificial disulfide bond as shown in figure 2.2. Correct folding of this mutant was indicated by circular dichroism (CD) experiments shown in figure 2.7A. AFM experiments yielded forces similar to those measured during the unfolding of the wild type spectrin repeat. However, as shown in figure 2.8A, the unfolding length was now typically around 14 nm, i.e. 18 nm shorter than the length gained from the complete unfolding of a native repeat. The wild-type pattern of unfolding lengths was re-established under reducing conditions (10 mM DTT) that disrupt disulfide bonds (figure 2.8B).

How exactly should the designed disulfide bond affect the unfolding lengths of the repeat? I derived a model for the double cysteine mutant from the water-refined NMR structure and calculated five unfolding trajectories, each covering 0.75 ns. The connection of helix A and B limited unfolding to a maximal exten-

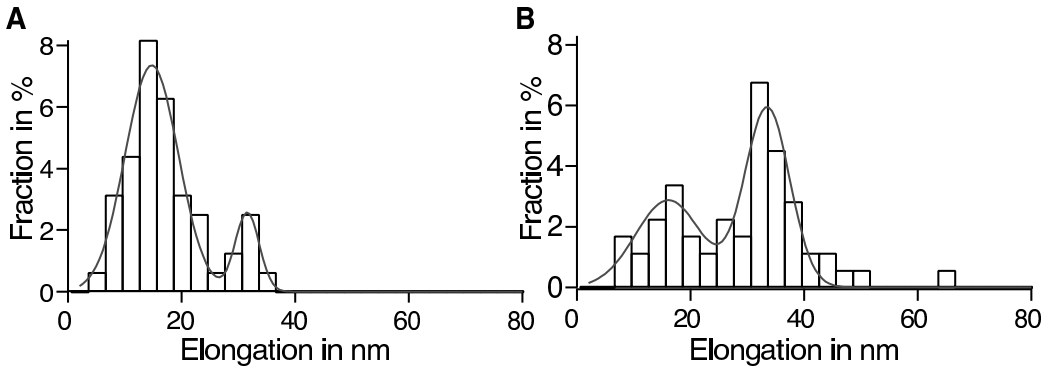

Figure 2.8: Forced unfolding of spectrin repeats containing a non-native disulfide bond. The distribution of elongation events is given for the unfolding of mutated tetrameric constructs in oxidized c) and reduced form d) (n=96). The lines in c) and d) each correspond to a fit by a normalized sum of two Gauss functions.

sion r_{NC} of about 19.5 nm. Employing the same conditions as before, I observed a linear force increase at extensions beyond 17 nm. The maximal extension, determined as described in section 2.2.2, was on average 18.1$\pm$0.1 nm r_{NC}, i.e. 17.7 nm shorter than in case of the wild-type repeat. The AFM experiments thus detected the expected shortening of the cross-linked double cysteine mutant with remarkable accuracy. This sub-nanometer precision indicated that the broad variation of particular unfolding lengths did not stem from experimental error.

The force profiles from simulations on the wild-type repeat (figure 2.4) suggest a rather different origin of the experimental variance. In simulations, initial force peaks were spread seemingly at random between elongations (from the NMR structure) of 1.5 up to 11 nm. This implies a random pre-stretching of the domain in AFM experiments. The precise unfolding length would be, in this scenario, obscured by the undefined pre-stretching of both the unfolding domain and of its replicas in the protein chain. However, what is the source of such remarkable "fuzziness"? The 11 pulling simulations were started from different time points of the same unperturbed deterministic simulation (without randomization of velocities) and were subject to identical conditions in all other respects. Their very different outcome in terms of pre-stretching and unfolding route is thus a pure effect of variations within the protein's native structure ensemble. This en-

semble effect puts thermal fluctuations into the position of a random generator for selecting the actual pre-stretching and force resistance of a particular domain.

2.3.4 Pathways and intermediates

The fluctuating topology of the simulated spectrin repeat depended mainly on the status of three hinge regions: loop AB, loop BC and a potential kink in helix B, and these three parameters provide a concise view on the trajectories (figure 2.4). One non-native force-resistant topology partly occurred at extensions which were still sustained by native-like folds in other simulations. It would hence further blur the distribution of unfolding lengths but should not show up as distinct unfolding event in AFM experiments.

However, other force-resistant topologies appeared at extensions that could well explain the additional experimental peak (snapshot e and f). Disruption of these topologies would result in length gains of 15 nm or more. All these force-resistant topologies (d, g, and e, f in figures 2.5 and 2.6 critically depended on the additional hinge inside helix B. To test the role of B helix bending in the unfolding pathways, my collaborators introduced two mutations that stabilized the central region of this helix. Like many spectrin repeats, R16 contains a proline and a glycine residue toward the middle of helix B. Yari Ylänne replaced this proline 62 and glycine 66 each by an alanine residue which increased the helix propensity at the putative hyphenation point. Indeed, the double alanine mutant proved thermally more stable than the native spectrin repeat, as is evident from the higher transition temperature during CD-monitored thermal unfolding shown in figure 2.7B. Pierre-François Lenne and Stephan Altmann studied the unfolding behavior of the mutant repeat with AFM.

The unfolding forces measured in the AFM were distributed broadly with a maximum between 40-70 pN (data not shown) and were thus again similar to those for the wild-type. However, significantly less curves displayed short peak-to-peak distances. The majority of peak-to-peak distances corresponded to complete unfolding events in the wild-type measurements (e.g. figure 2.3A). The resulting histogram of peak-to-peak distances is given in figure 2.9. The distribution of

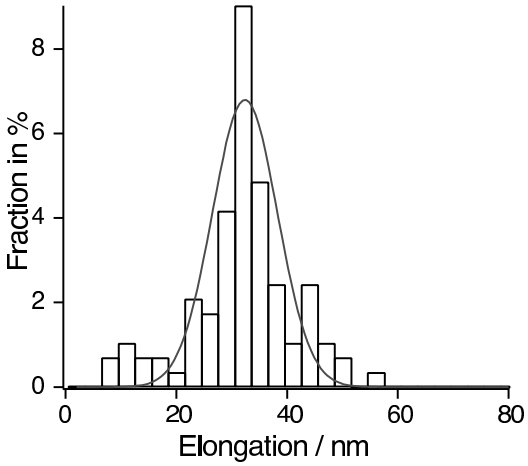

Figure 2.9: Probability distribution of elongations after unfolding events of double alanine spectrin constructs (n=120). The peak at 15 nm, present for the wild type in figure 2.3C has almost completely diminished. A single Gauss fit gives a maximum at 31 nm.

elongations peaks around 31 nm, similar to the second peak of the wild-type distribution in figure 2.3C. In contrast to the wild-type, the peak-to-peak distribution did not contain a statistically significant peak at 15 nm.

This result was corroborated by simulations. I modeled the double alanine mutant from the wild-type NMR structure and subjected it to 5 unfolding simulations employing identical conditions as before. No assumptions were made as to the influence of the mutations on the protein's backbone conformation. Nevertheless, during equilibration helix B straightened, and unfolding started, on average, with helix B angled at 134.5° (±3.8), compared to 122.4° (±4.9) in case of the native repeat. The maximum force observed during unfolding was on average 483±22 pN and thus similar to the native spectrin repeat. However, the trajectories exhibited fewer force peaks and, with one exception, all of them occurred before disruption of the native fold. In terms of topology, all trajectories closely followed either of the two main unfolding pathways which had already emerged from the wild-type simulations. However, figure 2.10A reveals that the relative importance of the two routes had reversed. In all but one trajectory loop BC was disrupted first and unfolding proceeded via the topology depicted by snapshot j

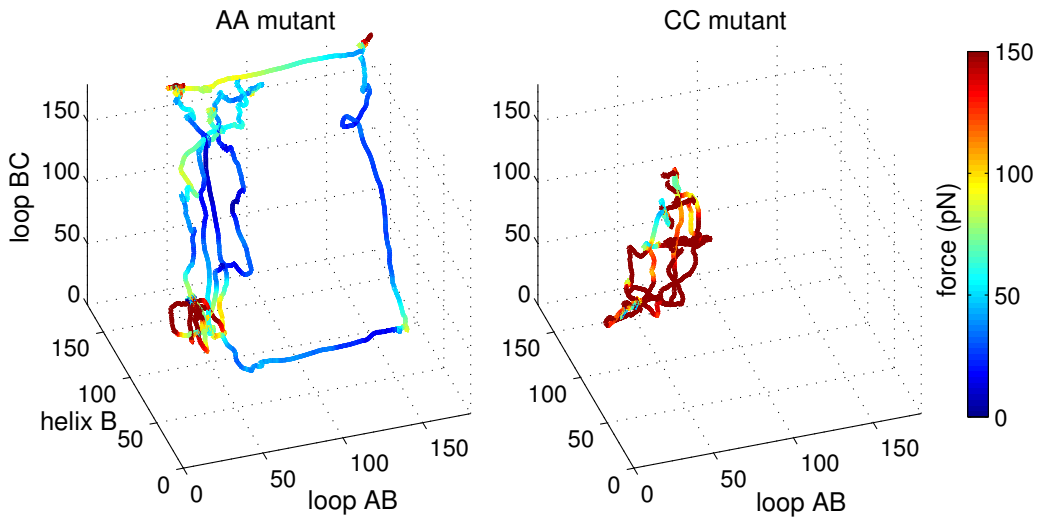

Figure 2.10: Unfolding pathways of mutated spectrin repeats. Hinge movements are defined in figure 2.2. Restoring forces are color-coded onto the smoothed angle traces as described in figure 2.5. Right: double alanine mutant, Left: double cysteine mutant. The hinge movement of helix B is restricted in the double alanine mutant and unfolding is steered away from the pathway that was most frequent in simulations of the wild type repeat.

in figures 2.5 and 2.6. In one case, such a topology caused high restoring forces around 22 nm r_{NC}, a possibility that I had not observed before. By contrast, wild-type simulations had shown two other regions of non-native topology with high resistance against unfolding. Only one trajectory of the double alanine mutant followed the formerly most important route and visited these topologies, which, however, unraveled without pronounced resistance. In summary, the two mutations in helix B appeared to disfavor structures that depended on the proposed hinge. Consequently, they steered unfolding away from the topologies that had produced all long forced intermediates in the wild-type simulations.

The outcome of experiments and simulations on the double alanine mutated spectrin repeat thus supports the proposed unfolding pathways with hinge-bending of helix B. Nevertheless, one can not yet exclude alternative roles of the replaced proline in a particular intermediate. However, a more general perturbation of the wild type would not explain our results since the mutations actually stabilized the native structure thermodynamically and did not alter the observed rup-

ture forces. Forced intermediates around an average extension of 15 nm are, in summary, supported by several observations: (1) AFM measurements of wild-type repeats featured a second prominent unfolding event around 15.5 nm gain of length. (2) Force resistant structures of appropriate length formed in MD simulations, thanks to a hyphenation point in the center of the molecule. (3) Two point mutations eliminated the short unfolding event by stiffening this hyphenation point. (4) More than four peaks were observed in some force curves, but within the expected maximum stretched length (compare figure 2.3B), and "half" unfolding events occurred mainly in earlier extension states at, on average, lower forces (Lenne et al. 2000).

2.4 Comparison with previous studies

Regular peak spacing and nearly constant rupture forces (200 pN at 1 m ms^{-1}) found in many studies on Ig-like domains from titin and fibronectin have established a standard for the type of results commonly expected from such experiments (Rief et al. 1997; Oberhauser et al. 1998; Carrion-Vazquez et al. 1999). In these studies, force curves were fitted to the worm-like chain model (Rief et al. 1997; Rief et al. 1998), implying a singular well defined rupture that transforms the native fold at once into a completely disordered protein chain held together by purely entropic effects. MD studies provided consistent mechanisms that convincingly supported such a defined all-or-none unfolding of domain I27 from titin (Marszalek et al. 1999). Similar predictions were initially made for FnIII domains although initial pre-stretching by 1-2 nm has now been suggested from more recent simulations (Craig et al. 2001) and two-step unfolding was observed in another simulation study (Paci and Karplus 1999).

Our experimental and theoretical results show that spectrin repeats react to an external force very differently than the Ig-like domains: (1) spectrin repeats can be stretched to varying extensions before the rupture of the triple-helical fold and (2) several unfolding pathways exist and some of them may lead to force-resistant non-native intermediate folds. For both reasons, the worm-like chain

model and, especially, the filtering of AFM results for regular peak spacing was not appropriate to describe the unfolding behavior. Experimental artifacts such as pickup of multiple molecules or surface interactions could be ruled out due to special experimental procedures described earlier (Lenne et al. 2000).

Two experimental as well as two computational studies previously examined forced unfolding of spectrin repeats, and on both sides there remained disagreement about whether two-step unfolding (i.e. forced intermediates) can occur or not. Rief et al. (1999) were the first to study unfolding of spectrin repeats with AFM. They used a hexameric construct of non-identical repeats and analyzed their data with the worm-like chain model. Concerned about pickup of multiple molecules, they only considered force curves with evenly spaced peaks. Consequently, they concluded that the repeat has a wider unfolding barrier than Ig-like domains but, nevertheless, unravels in a single step. By contrast, Lenne et al. (2000) studied the described tetrameric construct of identical repeats and used different criteria to filter out multiple pickups. They found both complete and half unfolding events and also reproduced this bi-modal distribution with Rief et al.'s hexameric construct.

On the theoretical side, Paci and Karplus (2000) subjected the spectrin repeat to MD unfolding simulations. Compared to our work, they used a different force field model, a rather different formulation of the pulling restraint, a faster unfolding regime, and a more approximate model of solvation. Nevertheless, their simulations seem to resemble our results in that they lack a singular well-defined force peak. The published average force profile also indicates high initial restoring forces distributed between 5 and ca. 15 nm r_{NC}. They suggested an intermediate to sometimes occur before this destruction of the original helix arrangement. Although I could not reproduce their suggested β-hairpin structure, I did observe similar force resistant structures at the described length. Snapshot c in figure 2.6 shows such a state that caused a near-maximum force peak at 13.5 nm r_{NC}. Paci and Karplus did not report later intermediates which, however, are necessary to explain the observation of peaks separated by about 15 nm.

Klimov and Thirumalai (2000) obtained predictions from a much more simpli-

fied computational model. They compared the native interactions between complete secondary structure blocks (in this case helices A, B and C) which they assumed to each unfold in all-or-none fashion. According to this model a spectrin repeat would unravel starting with helix A, without exhibiting any intermediates. They assumed that it is mainly the native topology that defines a protein's reaction to external force – a picture supported by previous simulations of forced (Lu and Schulten 1999; Paci and Karplus 2000) and thermal (Gsponer and Caflisch 2001) unfolding. Their pathways predicted for the unfolding of two Ig-like domains agree with MD simulations done on those systems. For spectrin, however, their predictions contradict our experiments and simulations. As a model is often most interesting when it fails, this one illustrates key findings of our collaborative effort: In case of spectrin, helix B can not be considered as a single structural block and forced intermediates can, in fact, arise from non-native topologies.

2.5 Conclusion

Spectrin is a mechanical protein. It has to withstand large deformations and responds elastically to external forces. Previous studies have shown that spectrin's structure is geared toward this function in many aspects. The topology of its global two-dimensional mesh work (Lee and Discher 2001), variable supercoiling of α and β subunits within the long tetramer (McGough 1999), and interactions between neighboring domains (Grum et al. 1999; MacDonald and Pozharski 2001) are apparently all contributing to elasticity. However, isolated spectrin domains react to external force in a manner that does not resemble the well defined and regular response commonly expected from "mechanical" protein domains. Hence, some authors (dis-)regarded the single repeat as "force compliant" (Paci and Karplus 2000; Carrion-Vazquez et al. 2000). By contrast, I suggest that this kind of behavior is yet another adaptation toward elasticity. Under mechanical stress the repeat choses between a variety of unfolding pathways. Each confers a different tolerance to forced elongation and some pathways proceed through non-native globular folds that may block unfolding even in the middle of com-

plete extension. This programmed fuzziness leads to an unusually large variation of unfolding lengths in AFM experiments and it multiplies with each additional spectrin repeat hooked to a chain of its *Doppelgängers.* It is certainly no coincidence that the domain mostly occurs in tandem with several replicas. A chain of many repeats should show nearly constant (if moderate) resistance to further unfolding at virtually any point along a very wide range of extension. Non-native intermediate states would increase the fuzziness gained per repeat and multiply the "working range" of this molecular spring, that is defer its "over stretching".

In my simulations such intermediate states arose from a hyphenation point in the central helix of the domain. Spectrin repeats often feature proline and glycine residues at this or neighboring positions and the helix is kinked in six out of seven known structures. The specific stiffening of this proposed hyphenation point did indeed abolish partial unfolding in AFM experiments. Moreover, forces measured by my collaborators are in the physiological range observed on erythrocyte ghosts (Sleep et al. 1999) and recent experiments on intact spectrin networks of erythrocytes tentatively suggested unfolding of single repeats (Lee and Discher 2001). Our combined data from single molecule experiments and MD simulations provide an intriguing glimpse on a molecule that may be optimized for elasticity on all scales of its architecture: from the structure of its cell-spanning network down to predetermined hyphenation points in its individual domains. At the level of the single spectrin repeat, elasticity appears to arise from a programmed diversity of unfolding pathways. Spectrin seems to translate thermal fluctuations of atomic structure into a smooth response to external force. The molecular dynamics of this protein is thus the link between structure and function.

2.6 Methods

2.6.1 Experimental methods

A detailed description of my collaborators' experiments is given in our joint publication (Altmann et al. 2002). The experiments were performed with repeat 16

from *Gallus gallus* non-erythroid α-spectrin (Wasenius et al. 1989), accession number P07751. I use amino acid residue numbering 1-116 for the chicken alpha spectrin residues 1762-1877, respectively.

2.6.2 Molecular dynamics simulations

Simulations were performed with the Amber 6.0 program package using the modified all-atom force field parm96 (Cornell et al. 1995; Kollman et al. 1997). Bond lengths involving hydrogen atoms were fixed with the SHAKE algorithm allowing for an integration time step of 2 fs. A cutoff of 15 Å was applied to non-bonded interactions. The temperature was controlled with the Berendsen coupling algorithm using a "coupling constant" of 5 ps. Effects of solvation were emulated with the generalized Born model and a tension term proportional to the molecule's surface area, both implemented in Amber 6.

The simulations were based on the solution structure of the 16^{th} repeat of chicken non-erythroid α-spectrin (Pascual et al. 1997) which had been subjected to an additional refinement in explicit water. The molecule was minimized and heated to 300 K with a linear temperature increase over 30 ps while atomic velocities were re-assigned from a Maxwell distribution every 2.5 ps. An equilibration for 250 ps was followed by the 1 ns production MD which yielded 11 restart files spaced 100 ps apart. Unfolding was initiated from these restart files by imposing a harmonic distance restraint (force constant 5.9616 kcal (mol Å)$^{-1}$) on the system which forced N and C termini apart at a constant velocity of 0.2 Å ps^{-1}. This approach was chosen for its ease of implementation and because the molecule can be assumed to align to an external force vector prior to any recorded unfolding events (Lu and Schulten 2000). Mutations were introduced into the water-refined structure without changes to the backbone geometry. In case of the double cysteine mutant, WhatIf (Vriend 1990) was used to suggest rotamers for the introduced cysteine side chains. The altered structures were subjected to the same protocol as the wild type but equilibration was extended to 500 ps before the generation of the first starting configuration for unfolding.

Trajectories were inspected with VMD (Humphrey et al. 1996). Angles and

distances were extracted with ptraj (included in the Amber package) and visualized in MatLab (The MathWorks, Inc.). Structure figures were prepared with Molscript (Kraulis 1991) using secondary structure assignments from Rasmol (Sayle and Milner-White 1995).

Chapter 3

The dynamics of protein-protein binding

3.1 Introduction

3.1.1 Networks of interacting proteins

Proteins hardly ever work alone. Over the last years, novel proteomic and genetic experiments (reviewed by Drewes and Bouwmeester (2003)) and sequence-based prediction methods (Valencia and Pazos 2002) have identified or suggested an increasing wealth of protein interactions. From the current data, it is estimated that there are in the order of 20.000 connections among the about 6300 yeast proteins (Bader and Hogue 2002). On average, each protein thus seems to team up with 6 partners. The situation becomes likely even more complex in the human proteome with 30.000 genes and a higher prevalence of alternative splicing (Figeys 2003). Yet, even for yeast, the available interaction maps are still far from complete and contain sizable rates of false positives (Bader and Hogue 2002). Furthermore, the data usually don't tell anything about the spatial arrangement of the two proteins, not even whether they are indeed in direct contact or only belong to the same macromolecular assembly.

On the other hand, protein complexes are difficult targets both for X-ray crys-

tallography and NMR experiments. Systematic structure determination efforts mostly concentrate on single proteins or domains (Zhang and Kim 2003). We are thus quickly accumulating data about large interaction networks and are also increasingly able to assign atomic structures to many of their components. Yet, unfortunately, we usually cannot piece the puzzle together – the structure of protein assemblies remains elusive. Even though selected protein complexes and their structures have been studied since decades, it seems, we haven't sufficiently understood the process of protein-protein binding.

Static properties of protein complexes, such as size (Lo Conte et al. 1999), amino acid composition (Jones et al. 2000) and shape of interfaces (Jones and Thornton 1997) have already been scrutinized in quite detail. However, little is known about the interplay between protein-protein binding and dynamics, except the fact that structures differ between free and bound state (Betts and Sternberg 1999; Lo Conte et al. 1999). Ironically, it was experiments on binding (of small ligands to selected proteins) that first established an effect of structure dynamics on protein function in general (Kirschner et al. 1966; Austin et al. 1975; Lancet and Pecht 1976). Thus, the proper treatment of molecular dynamics may well be the missing link in our description of the binding process.

3.1.2 Current models of protein recognition

Our current understanding of protein recognition is caught in a contradiction: On the one hand experimental rates of association suggest that, in many cases, almost every collision between two partner proteins leads to the formation of a complex (Northrup and Erickson 1992). On the other hand, even if we know the atomic structure of both proteins, we often fail to predict the structure of the complex because the free partners simply do not fit sufficiently well. Over the last two decades the computational solution of this protein-protein docking problem has been an area of intense research (reviewed by Halperin et al. (2002)). Advances in docking methods often went hand in hand with new insights into the binding mechanism.

Structures of protein complexes reveal intricate shape complementarity be-

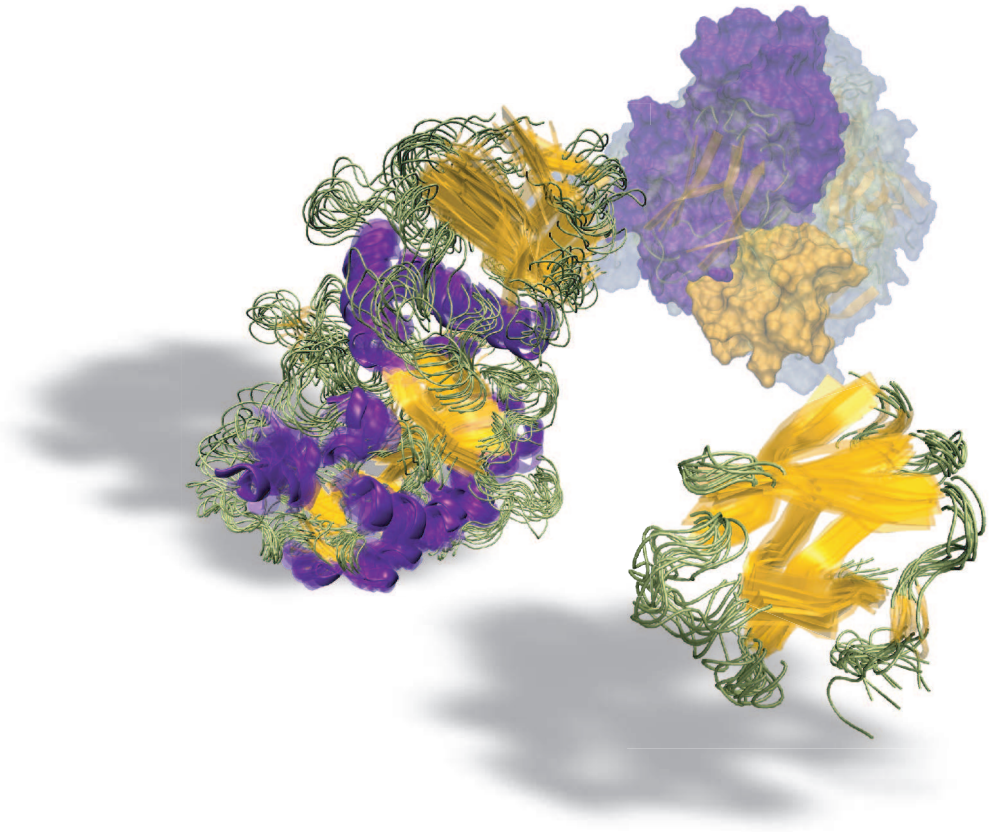

Figure 3.1: How can two proteins recognize each other if they don't fit in first place? The extent of protein flexibility is illustrated by two structure ensembles obtained from (PCR-MD) simulations of the enzyme glycosyltransferase (left) and its small inhibitor tendamistat (front right). The native arrangement of the enzyme (purple) and inhibitor (golden) in their complex hovers as a solid surface model in the background and is surrounded by several alternative but wrong orientations of the ligand.

tween the binding partners, which seemingly confirms Emil Fischer's (1894) key-lock model of biomolecular interaction. However, the free (unbound) receptor and ligand structures are often much less complementary and show significant deviations from their bound conformation (Betts and Sternberg 1999; Lo Conte et al. 1999). Consequently, early rigid-body docking algorithms could re-dock known complexes but were unable to predict them from the free components (Kuntz et al. 1982; Goodford 1985). The key-lock model may hold for the final protein complex but it cannot explain the process of recognition between the free molecules.

Daniel Koshland's (1958) induced fit model acknowledges a certain plasticity of proteins and postulates a mutual adaptation of the two structures. It offers a valid description of recognition if we assume that this process is driven by forces that do not require good shape complementarity to start with. However, protein-protein recognition seems to be controlled, to a large extent, by short range electrostatics (Frisch et al. 2001), desolvation entropy (Camacho et al. 2000), and van der Waals interactions (Gray et al. 2003), which all depend to various degrees on shape complementarity. Induced fit may be appropriate for describing the transformation of receptor and ligand after recognition has occurred, but it cannot explain the process of recognition itself (Bosshard 2001).

A third model, conformational selection, is inspired by the MWC mechanism of allosteric regulation (Monod et al. 1965) and is more compatible with short range interaction forces. Experimental protein structures are only the average of many conformational states (Frauenfelder et al. 1991). The model postulates "recognition" conformers that are hidden in the two structure ensembles and select each other upon binding. Early on, experiments corroborated the MWC model (Kirschner et al. 1966). Later experiments on antibodies showed that, in several cases, binding of an antigen was influenced by an equilibrium of different antibody conformations (e.g. Lancet and Pecht (1976),Foote and Milstein (1994)). Experimental evidence was also provided for the inverse case – the selection of antigen conformers by antibodies (Leder et al. 1995; Berger et al. 1999). Kumar et al. (2000) then suggested conformational selection as a mechanism for protein-protein interaction in general. They explicitly postulated that bound con-

formations of receptor and ligand are part of their free structure ensembles and that recognition occurs between the two bound conformers. Thus, recognition and (apparent) structural adaptation could be explained simultaneously. Evidence for a preexisting equilibrium between free and bound conformations is hard to come by. Recent experimental structures are interpreted in this direction (Goh et al. 2004). However, at closer examination they confirm the existence of distinct conformations in free and bound structure ensembles but only very few suggest overlaps between the two. Since it usually leaves no traces in free crystallographic or NMR structures, the bound conformation, if it is present, must be a rare state.

3.1.3 The kinetics of interaction

The elegance of the preexisting equilibrium hypothesis stems from its combination of the modern ensemble view of protein structure with a simple key-lock mechanism for recognition. However, the model is challenged by the usually very fast pace of protein-protein recognition, which does not leave room for many unsuccessful collisions (Northrup and Erickson 1992). Recognition conformations must be frequent enough to occur simultaneously for both receptor and ligand within the short time window during which they are properly aligned in the course of a single random collision. Northrup and Erickson describe a protein encounter as a series of micro collisions at different orientations. Estimates for the length of a (possibly correctly) aligned micro collision range from 400 ps as lower bound to 10 ns as upper bound (Northrup and Erickson 1992; Janin 1997). The preexisting equilibrium hypothesis thus implies a certain minimum frequency of bound conformations. After all, this conformation has to occur simultaneously in both the receptor and ligand ensemble during the short time of random encounter. A rough calculation highlights this problem:

The recognition probability R of a correctly aligned micro collision should depend on the average frequencies $\langle fr \rangle$ of recognition conformations in the two free ensembles. The probability of recognition failure can be estimated as:

$$1 - R = \left(1 - \langle fr \rangle^2\right)^N , \tag{3.1}$$

where N is the number of distinct conformations sampled in the course of the correct alignment. The frequency of recognition conformations which is needed for a certain recognition rate is then

$$\langle fr \rangle = \sqrt{1 - \exp \frac{\ln(1-R)}{N}} \,. \tag{3.2}$$

N depends on the lifetime τ of the alignment and on our definition of distinct conformations. The short recognition time will only allow for fairly limited sampling in the flat energy landscape of protein structures. For the sake of simplicity, I assume that N depends linearly on the recognition time τ and that the "recognizability" of a given protein structure changes every 1 ps ($N = \tau/\mathrm{ps}$). According to this rough estimate, bound conformations must represent 4% of both free ensembles in order to achieve a 50% recognition success within a 400 ps time window. Even a fairly unrealistic recognition time of 10 ns still requires a frequency close to 1%.

A valid model of protein-protein association needs to explain not only the obvious difference between free and bound protein structures, but must also be compatible with kinetic data. So far, the two problems are usually addressed in isolation. The detailed theoretical studies on the kinetic mechanism of binding have focused on the diffusion of proteins that are rigidly locked into their bound conformation (Northrup and Erickson 1992; Janin 1997; Camacho et al. 1999; Selzer and Schreiber 2001; Zhou 2001). These models can reproduce the kinetics of diffusion-controlled protein-protein associations with some success (Gabdoulline and Wade 2002) but regard structural transitions only as a passive induced fit after recognition has occurred.

Likewise, protein-protein docking algorithms rely on rigid body, rigid segment (Schneidman-Duhovny et al. 2003) or rigid backbone simplifications and regularly fail in the face of backbone motions (Gray et al. 2003). The effective treatment of overall protein flexibility is now the largest obstacle both to our understanding and to the reliable prediction of protein-protein association.

3.1.4 The thermodynamics of interaction

Fixing one binding partner to the other evidently comes with a significant cost of (translational and rotational) entropy. For two proteins, this entropy loss should amount to about 100 $\mathrm{cal\,mol^{-1}K^{-1}}$ (Janin 1995). From the analysis of crystalline proteins, Finkelstein and Janin (1989) estimated that it is in part compensated by residual motions of the two partners within the complex. Such motions should contribute in the order of 50 $\mathrm{cal\,mol^{-1}K^{-1}}$ and improve the overall entropy balance from -100 to about -50 $\mathrm{cal\,mol^{-1}K^{-1}}$. However, this still corresponds to a free energy difference of 15 $\mathrm{kcal\,mol^{-1}}$ in favor of dissociation.

Intuitively, binding is usually assumed to also restrict the flexibility of both proteins and, as a consequence, to claim an additional cost of conformational entropy. Both the idea of induced fit and the preexisting equilibrium model imply that the complex has less conformational freedom than the unbound components. However, the extent of this entropy loss, or whether it is a loss at all, remains controversial.

Computational studies often estimate the conformational entropy loss from the restriction of side chain rotameric states (e.g. Janin (1995)). By contrast, Tidor and Karplus (1994) concluded from normal mode analysis that the dimerization of insulin is actually promoted by an 23 $\mathrm{cal\,mol^{-1}K^{-1}}$ increase of conformational entropy. Subsequent calculations on different complexes using different methods generally estimated an overall conformational entropy loss (Viñals et al. 2002; Gohlke and Case 2004; Hsu et al. 2004) but gains were reported for one component of a complex (Hsu et al. 2004).

Experimental studies arrived at mixed results. Thermodynamic experiments such as microcalorimetry cannot separate changes of protein conformational entropy from the entropic contributions of the solvent. However, NMR relaxation studies are able to measure the fluctuations of selected backbone and side chain atoms (see 1.2.2 on page 5). Several groups estimated changes of conformational entropy from such (incomplete) data. A significant loss of conformational entropy was derived for the interaction of specific peptides with calmodulin (Lee et al. 2000), troponin C (Mercier et al. 2001), and the c-Src SH3 domain (Wang et al.

2001). However, the converse was found for other interactions. The binding of a small ligand to mouse major urinary protein (Zidek et al. 1999), the association between inhibitor TIMP-1 and the catalytic domain of stromelysin 1 (Arumugam et al. 2003), as well as the recognition of an anchoring protein by the D/D domain of cyclic-AMP dependent protein kinase (PKA) (Fayos et al. 2003), all appear to proceed with a substantial gain of conformational entropy. Forman-Kay (1999) summarized the data available at that time and concluded that upon binding motions can increase, decrease or stay the same. She suggested that increased motion may in some cases deliver a critical supplement to the free energy of binding.

Protein motions clearly have considerable influence on the stability of protein complexes. However, calculations and experiments find no overall trend even for the sign of this contribution. The common assumption, that binding generally restricts flexibility and occurs at the expense of conformational entropy, appears not justified by current data. Yet, both calculations and experiments suffer from some serious shortcomings. The NMR studies cited, with the exception of Lee et al. (2000), based their estimates solely on the measurement of backbone amide fluctuations. Computational approaches consider the whole protein but rely on other critical simplifications. Normal mode calculations ignore any anharmonic motions of the protein. The alternative analysis of molecular dynamics simulations encounters convergence problems due to insufficient sampling (Gohlke and Case 2004; Hsu et al. 2004). Furthermore, the available experimental and theoretical studies focussed each on a single interaction. A consistent picture of how the thermodynamics of binding is influenced by protein motions has yet to emerge.

3.1.5 Our approach

In this chapter I examine the interplay of overall protein flexibility and protein-protein binding (figure 3.1). I selected a set of 17 protein complexes for which the three-dimensional structures of both free components and the complex are available. For each of these 51 molecules I performed short (10 x 50 ps) molecular dynamics simulations in explicit water. I find that uncomplexed binding interfaces are more flexible than the remaining surface and that they loose conformational

freedom upon complex formation. Nevertheless, in the majority of cases binding does not restrict the overall motion of the proteins. I calculated the change in conformational entropy from longer simulations (10 x 1 ns) on the free and bound state of 7 complexes. Two small complexes and an antibody-antigen system exhibited a significant loss, whereas three larger complexes showed increased or unchanged conformational entropy.

I then combined the molecular dynamics based sampling with systematic rigid body docking. I applied shape-driven docking to all combinations of representative snapshots from the free structure ensembles of the 17 receptor and 16 ligand proteins. I compared the success of this extended but still manageable search with the simple docking of the experimental structures. Already very sparse structure ensembles contained several combinations of receptor and ligand conformers that generated more and better near-native solutions. Remarkably, the docking performance of a given combination of receptor and ligand structure was largely uncorrelated with their similarity to the bound conformation. Based on these results, I extend and combine the up to now conflicting models of protein-protein binding. I suggest a 3-step mechanism of diffusion, free conformer selection and refolding as working model for flexible recognition.

Most of this work results from a close collaboration with Johan Leckner, who performed his postdoctoral research in our lab. For many parts, it is difficult to discern his from my contributions. Johan Leckner compiled the set of 17 complexes and benchmarked various docking programs. In general, he was more responsible for the docking aspects of this project, whereas I focused more on the molecular dynamics simulations. The analysis of flexibilities was largely, the study of entropies entirely my assignment. Together, we built up a library of programs that became instrumental for the automation of calculations as well as their analysis and visualization. Two other colleagues, Michael Habeck and Wolfgang Rieping, contributed algorithms to this library. They also helped with mathematical problems, especially the statistical formulation of docking specificity (section 3.4.5). Moreover, Michael Nilges provided protocols for X-Plor simulations and helped with their implementation. Many of the following sections are kept in ”we” narrative as they present collaborative work. In most cases, this ”we” should be

translated to ”Johan Leckner and I”.

I divided this chapter into three parts. First, I examine the dynamics encountered *before* binding, that is to which extent the flexibility of free binding patches differs from the remaining surface. Second, I focus on the situation *after* binding has occurred; I compare the flexibility of free and bound state and, after resolving some technical issues, estimate the entropic contribution to the stability of several complexes. The last part examines the process of recognition itself. It is a repetition of the joint publication with Johan Leckner (Grünberg et al. 2004). The figures of this section were mostly prepared by Johan Leckner; The original text, including the proposed model, was mostly written by myself.

3.2 The flexibility of free binding interfaces

3.2.1 Current notions of flexibility

Sundberg and Mariuzza (2000) concluded from a review of experimental studies, that increased flexibility may have advantages, in particular, for proteins that need to recognize different ligands at a single binding site. Halperin et al. (2002) extended this argument. They predicted higher flexibilities also for ”normal” binding sites, as a way to facilitate recognition and better adapt to mutations in the interaction partners. Luque and Freire (2000) used unfolding simulations to analyze the stability of 16 proteins with binding sites for small molecules. They found that binding sites comprised regions of low next to regions of high structural stability. Such a dual character with high and low flexibility was also suggested by Ma et al. (2003) from the study of residue conservation. In contrast, Cole and Warwicker (2002) examined the flexibility of side chains in protein-protein interfaces (the rotamers available after separating the complex) and concluded that binding patches were generally less flexible than the remaining surface. They expected that a reduced flexibility would limit the entropic cost of binding.

Hence, free binding interfaces have been suggested to be more flexible, more rigid or of dual character. All three variants can be motivated with certain models

of recognition in mind. The cited predictions were based on the study of bound structures (taken from the complex) with different approximate models. In the following, we subject a set of 33 uncomplexed proteins to molecular dynamics simulations in explicit water. All of these proteins are involved in transient interactions, that is they exist independently, but we also know their structure in a protein-protein complex. We can hence identify the free binding interface and compare its dynamics to the remaining surface.

3.2.2 Structural data

We selected a set of 17 protein-protein complexes for which the structures of both the free components and the complex are available. Table 3.1 gives a description of these complexes. The set is based on docking benchmarks from Graham Smith (http://www.bmm.icnet.uk/docking/systems.html) and Chen et al. (2003). From these benchmarks, we excluded complexes with large non-protein ligands to facilitate the mostly automated modeling procedure. Only the free structures and molecular dynamics ensembles derived from them were used for the analysis of surface flexibility. The structure of receptor and ligand solved as a complex served for the definition of the binding patch.

3.2.3 Conformational sampling

Rather than by a static structure, proteins are best described by an ensemble of individual conformations (Frauenfelder et al. 1991). This section examines protein flexibility before the onset of binding and thus concentrates on conformational ensembles of free receptors and free ligands. Molecular dynamics (MD) simulations offer a way to generate such ensembles (Frauenfelder and Leeson 1998). However, even long and computationally expensive simulations cannot insure complete sampling. We performed simulations on 33 different proteins (receptors and ligands; c06 and c08 share one ligand) and later extended the analysis to the 17 bound ensembles (see section 3.3). The calculation of 50 ensembles requires a compromise between computational cost, sampling coverage and accuracy.

Table 3.1: Protein-protein complexes examinded in this study.

ID[1]	Receptor / Ligand	PDB codes, chain identifier			Residues	
		rec	lig	com	rec	lig
c01	Trypsin / Amyloid β-protein precursor inhibitor domain	1BRA	1AAP A	1BRC E:I	223	56
c02	α-chymotrypsinogen / Pancreatic secretory trypsin inhibitor	2CGA A	1HPT	1CGI E:I	245	56
c03	Kallikrein A / Pancreatic trypsin inhibitor	2PKA AB	5PTI	2KAI AB:I	232	58
c04	Kallikrein A / Pancreatic trypsin inhibitor	2PKA AB	5PTI	2KAI AB:I	232	58
c04	Subtilisin BPN / Subtilisin inhibitor	1SUP	3SSI	2SIC E:I	275	108
c05	Tissue factor extracellular domain / Antibody Fab 5G9	1FGN LH	1BOY	1AHW AB:C	248	211
c06	Humanized anti-lysozyme Fv / Lysozyme	1BVL AB	3LTZ	1BVK AB:C	224	129
c08	Anti-lysozyme antibody Hyhel-63 / Lysozyme	1DQQ AB	3LTZ	1DQJ AB:C	424	129
c11	Barnase / Barstar	1A19 A	1A2P A	1BSG A:E	108	89
c13	Ribonuclease inhibitor / Ribonuclease A	2BNH	7RSA	1DFJ E:I	456	124
c14	Acetylcholinesterase / Fasciculin-II	1VXR	1FSC A	1FSS A:B	532	61
c15	HIVB-1 NEF / FYN tyrosin kinase SH3 domain	1AVV	1SHF A	1AVZ B:C	99	59
c16	Uracil-DNA glycosylase / Inhibitor	1AKZ	1UGI A	1UGH E:I	223	83
c17	RAS activating domain / RAS	1WER	5P21	1WQ1 R:G	324	166
c19	Glycosyltransferase / Tendamistat	1PIF	2AIT 1	1BVM P:T	495	74
c20	CDK2 cyclin-dependant kinase 2 / Cyclin A	1HCL	1VIN	1FIN A:B	294	252
c21	CDK2 cyclin-dependant kinase 2 / KAP	1B39 A	1FPZ A	1FQ1 A:B	290	176
c22	Heteromeric G-protein / Transductin Gt-α	1TBG AE	1TAG	1GOT A:BG	408	314

[1]Complex identifier used throughout the chapter (retained from www.bmm.icnet.uk/docking-/systems.html)

For each protein, we calculated 10 independent trajectories of 50 ps length each, with the structure embedded in a 9 Å layer of explicit water. The use of multiple short instead of a single long trajectory is expected to increase sampling by a factor of 2 (Caves et al. 1998). Besides, protein-protein recognition is assumed to occur in a short time window of about 400 ps (Northrup and Erickson 1992; Janin 1997). Our simulations should thus cover motions on the time scale relevant to recognition. For simulations on spectrin (chapter 2), I had used an implicit solvent model to reduce computational cost and to alleviate artifacts of the high unfolding speed. However, for the study of protein recognition, we depended on a realistic representation of surface dynamics. The incorporation of explicit water was therefore important and the Langevin regime constituted a good compromise between expense and accuracy (Izaguirre et al. 2001). Several of the comparatively short simulations were backed up by longer (10 x 1 ns) simulations using a different force field and periodic boundary conditions. A description of these calculations follows in section 3.3.1.

3.2.4 Definition of flexibility

Flexibility can be measured in several ways. In the following, I define flexibility as the average distance between the snapshots of a conformation ensemble. The snapshots were taken in regular intervals from the second half of mutually independent simulations, that is I did not compare snapshots stemming from the same trajectory (see 3.6.4 for details). The mean of the distribution of pairwise distances characterizes the diversity (flexibility) of a structure ensemble. The width of the distribution (which I denote ”spread” of diversity) describes the range of similarities and would indicate if the mean was distorted by distinct subpopulations of closely related structures. The evaluation of pairwise distances eliminates the need to choose an arbitrary reference structure for the whole ensemble. The focus on snapshots from independent simulations diminishes the influence of the sampling interval. Otherwise, the generally higher similarity of neighboring snapshots would create a subpopulation of small distances and distort mean as well as width of the overall distribution.

A serious drawback of root mean square (rms) distances is their dependency on the size and shape of a structure. It would be misleading to simply compare rms distances between snapshots of the binding patch with rms distances of snapshots from the whole surface. For this reason, I divided the protein surface into random patches, each having the same number of atoms as the binding interface. These random patches provide the appropriate reference flexibility along with an estimate of its common variation across the non-binding surface of a given protein.

3.2.5 Surface flexibility

Free binding interfaces are more flexible than the remaining surface of the protein. Figure 3.2 compares the flexibility of binding interfaces with the mobility of random surface patches from the same molecule. At least on the short time scale simulated, the ensembles of binding patches are always more diverse (gray bars) than the conformations of the average non-binding patch (solid line). In most cases, the difference surpasses also the standard deviation of random patch flexibilities (error bars) and is thus significant even for the single protein.

Studies of protein-protein recognition often pay special attention to the dynamics of amino acid side chains in the context of a rigid protein backbone (Najmanovich et al. 2000; Kimura et al. 2001; Cole and Warwicker 2002; Rajamani et al. 2004). However, the distinction between backbone and side chain dynamics is dictated by technical constraints and lacks a physical basis. Side chain and backbone torsions are correlated (Schrauber et al. 1993). Upon binding, side chain and backbone atoms are equally involved in conformational changes (Betts and Sternberg 1999; Lo Conte et al. 1999). Furthermore, also backbone conformations display significant variations across independently determined structures (Chothia and Lesk 1986) and deformations on this scale can already affect docking results (Ehrlich et al. 2005). From this point of view, such a thing as side chain flexibility does, strictly speaking, not exist.

This view is confirmed by the analysis of surface flexibility. As shown in figure 3.3 the higher mobility of binding interfaces extends well to the protein backbone

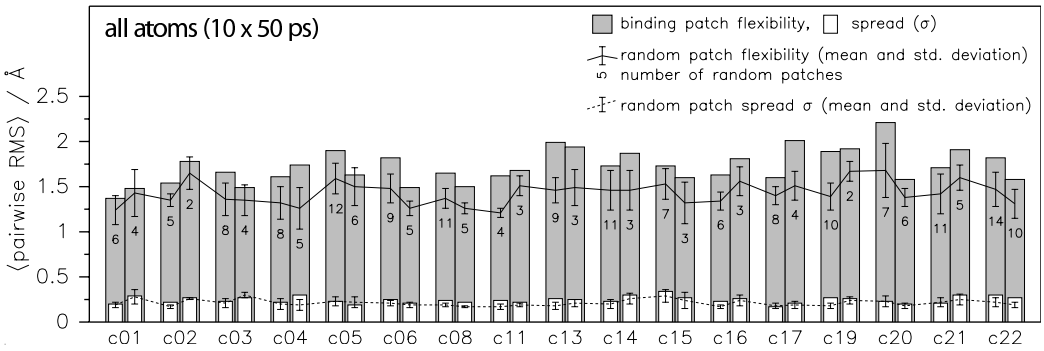

Figure 3.2: Free binding interfaces are more flexible than random surface patches. Each complex is represented by a bar for receptor (left) and ligand (right), respectively. The flexibility of binding patch snapshots (gray bars) is compared to the average flexibility (average of average pairwise rmsd) of random surface patches with the same number of atoms (solid line). Error bars describe the confidence of this average flexibility (+ and - 1 std. dev.). Open bars quantify the spread (standard deviation) of pairwise distances; the associated broken line and error bars indicate the spread averaged over the random patches.

(C_β atoms were included because their motion is tightly linked to backbone fluctuations). Thus, the high flexibility cannot be attributed to a special diversity of side chain conformations but constitutes a property of the whole interface. Residues of particular functional importance quite definitely exist (Yao et al. 2003) and their dynamics is worth studying. However, the artificial separation of side chain and backbone motions is certainly not improving the description of protein-protein recognition.

For 7 complexes and their components, I performed additional simulations of 10 x 1 ns length with a more elaborate treatment of solvation and electrostatics. The flexibility of binding patches is significantly increased also in most of the longer simulations. Some differences emerge in detail. In particular, the binding patch of barstar (c11 ligand) turns out more rigid when compared to the remaining surface. Interestingly, the other differences between the two simulation regimes diminish, if I only consider shorter sections from the beginning of the long trajectories (data not shown). The high mobility of the binding patch might be more pronounced on the short time scale relevant to recognition than on longer time scales (where it could impair the thermodynamic stability of the complex).

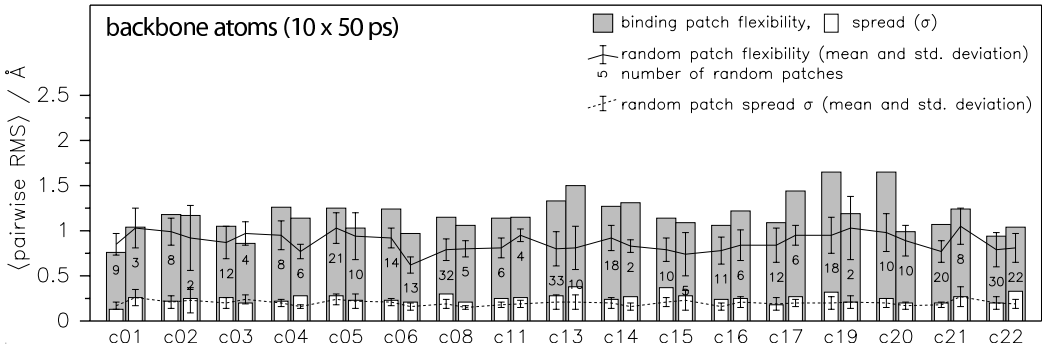

Figure 3.3: The higher mobility of binding interfaces also holds for the protein backbone. See figure 3.2 for a detailed description. In difference to figure 3.2, pairwise distances were calculated only between C, N, O, C_α, and C_β atoms of binding and random patches.

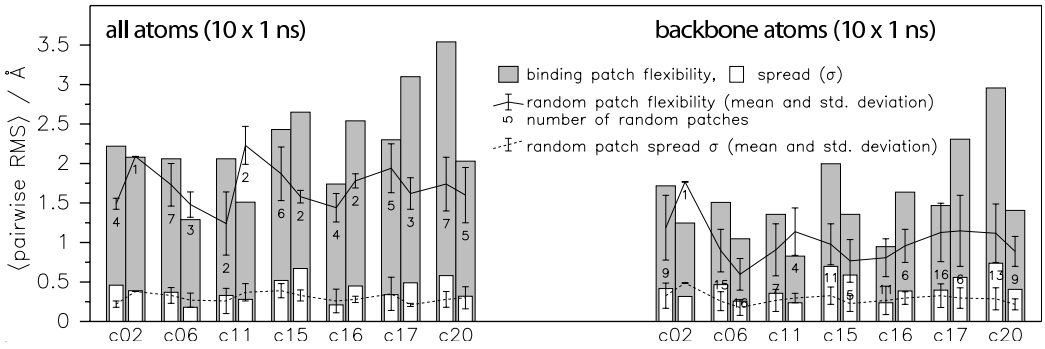

Figure 3.4: Surface flexibilities on a longer time scale. The mobility of binding interfaces and random surface patches is determined from the second half of more elaborate simulations over 10 x 1 ns. The representation is the same as in figure 3.2. The trend of higher binding patch mobility is confirmed although differences emerge in detail (see text).

The spread (range) of diversity shows no distinctive features of binding interfaces. It mostly scales with the overall higher mobility. Spacious "key side chains" in interaction interfaces were suggested to preferably occupy very few mutually distinct subconformations (Rajamani et al. 2004). The existence of pronounced subpopulations in a structure ensemble should manifest itself as a combination of low or average flexibility and large spread. In other words, the normally bell shaped distribution of pairwise distances should become bimodal or otherwise broadened. Such a trend is not evident. At least on the global scale examined here, binding interfaces show no particular tendency to "jump" between distinct conformations.

A dual character of binding interfaces with more rigid and more flexible parts existing side by side, is more difficult to determine and would require an in depth analysis of sub-patches or residue-centered motions. The overall increased mobility of the whole binding region presumably complicates the detection of any such trend.

3.3 Free and bound structure ensembles

3.3.1 Extended conformational sampling

For 21 molecules (7 complexes and their free components) the short simulations described above were backed up by 10 x 1 ns simulations in a solvent box using periodic boundary conditions, particle mesh Ewald treatment of electrostatic forces (Essmann et al. 1995) and a different force field (Cornell et al. 1995; Kollman et al. 1997).

The overall diversity of the different structure ensembles correlated well between the two very different simulation setups, besides the fact that it was of course generally higher in the longer simulations. In the previous section, I had quantified diversity or flexibility by the average of pairwise distances between the members of a structure ensemble. I had also introduced a measure for the "spread" of similarity, which was simply the standard deviation of the pairwise

distances. I compared the flexibility calculated from the last 10 x 30 ps of the short simulations with the same value calculated from the last 10 x 500 ps of the long trajectories. The correlation was best for the diversity and spread (R=0.95 and R=0.97, respectively) of the free ensembles, albeit after the removal of one outlier (c17 receptor). The two simulation regimes agreed somewhat less on the flexibility and spread of the 7 protein complexes. The flexibility correlated with R=0.77, the spread of diversity with R=0.88, and the latter value again excludes one outlier (c15 complex).

The short simulations did not adequately sample the slow residual intermolecular motions of receptor and ligand in protein complexes (which will be discussed in section 3.3.6) and did not sufficiently converge for the calculation of entropies. Nevertheless, the structural flexibility correlated surprisingly well between the two setups. The less elaborate simulations are thus sufficient for the study of this property.

3.3.2 Flexibility before and after binding

The highly mobile binding interfaces loose their conformational freedom upon formation of the protein complex. This is shown in figure 3.5, which presents the average over the flexibilities of all 33 proteins. Outside the contact region, surface atoms often experienced moderate gains of mobility (this is not an artifact of superpositioning, also the conformations of the complex were fitted separately for receptor and ligand). Surprisingly, the overall flexibility of a protein could both rise or fall; There was no general trend in either direction. The common assumption that binding restricts the flexibility of proteins, seems, in this generality, to be wrong. This absence of a trend would confirm the picture emerging from experimental studies (Forman-Kay 1999).

However, our set of 17 protein complexes is probably not representative. In fact, it appears difficult to define any ”representative” set of complexes. A subset of the 17 protein complexes, showed the same overall (absence of a) trend in the longer, more elaborate simulations (figure 3.6). The 7 protein complexes were not selected to mirror the flexibility of the whole set. They thus are a non-

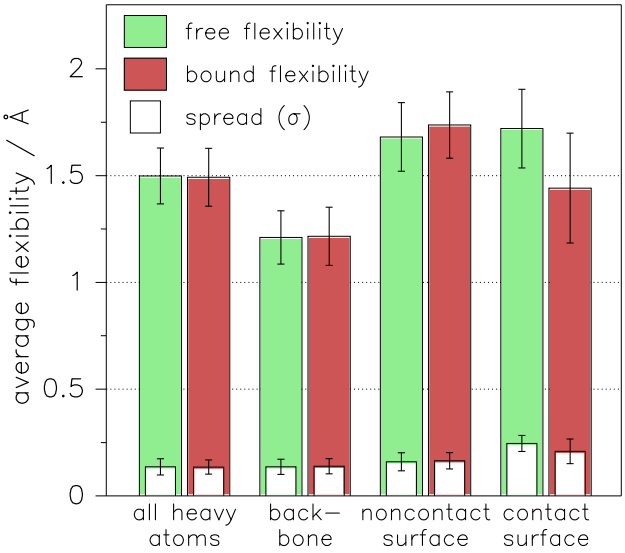

Figure 3.5: The flexibility of 33 proteins before and after binding. The values are averaged over all 33 proteins (receptors and ligands of 17 complexes). Binding surfaces generally loose, the non-contact surface often gains flexibility. The flexibility of the overall protein appears, on average, unaffected by binding. Note, that flexibilities cannot be compared between the different protein parts, because the rmsd measure also depends on size and shape of the selected region.

representative subset of a non-representative selection.

In general, the formation of transient protein-protein complexes severely restricts the flexibility of the binding interface. However, other parts of the protein appear often to compensate for this loss of mobility. This could, in fact, should have consequences for the entropy cost or gain of binding. How does protein flexibility influence the stability of the complex? Unfortunately, out of the different components of free energy, conformational entropy is currently the most difficult one to calculate (Gohlke and Case 2004). The following three sections address some of the technical problems involved. Sections 3.3.6 and 3.3.7 then extract estimates from the simulation data. As it turns out, conformational entropy has a considerable impact on the thermodynamics of protein-protein binding.

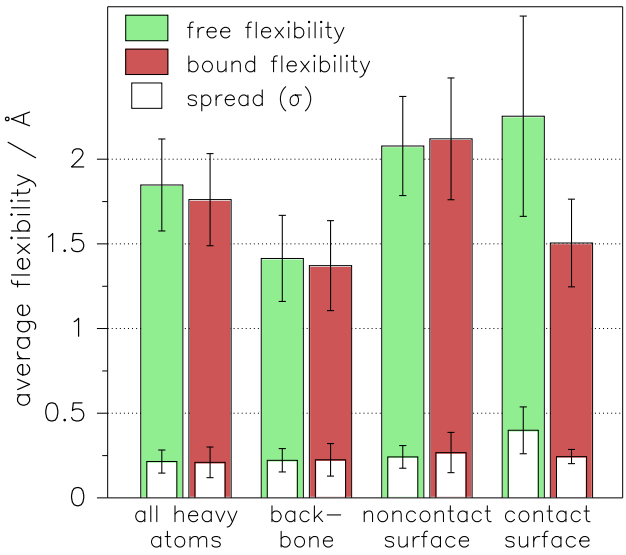

Figure 3.6: The flexibility of 14 proteins, simulated on a longer time scale, before and after binding. The set of 14 proteins is not representative for the set of 33 proteins in figure 3.5.

3.3.3 Quasiharmonic analysis and conformational entropy

The conformational entropy of proteins is commonly estimated from normal mode calculations on single structures in gas phase. The quasiharmonic analysis of molecular dynamics simulations offers a promising alternative, as it incorporates effects of anharmonic motion and solvation (Teeter and Case 1990). The method was briefly introduced in section 1.3.5 on page 15. So far, two simulation studies have attempted to apply quasiharmonic procedures to the analysis of protein-protein interaction. Both groups effectively discarded the approach.

Hsu et al. (2004) examined the binding of HIV-1 protein gp120 to its receptor CD4. They performed three 10 ns simulation and used an heuristic formula (Schlitter 1993) to estimate conformational entropies. However, they did not calculate the entropy of the complex but only compared the diversity of the individual components in free and bound state. Entropies did not converge over the 10 ns simulation. Therefore, they derived entropy differences from the heuristic concatenation of free and bound trajectory segments (along the time axis). By adding

up the individual change of gp120 and CD4, they arrived at an overall change of conformational entropy that agreed with an experimental measurement. However, as they note, the experimental value contains the contribution from desolvation, which was not at all considered in the calculation. This contribution should be large and the agreement is thus rather curious.

Gohlke and Case (2004) studied the interaction between H-Ras and the Ras-binding domain of C-Raf1 by three 12 ns simulations. They employed the more strictly "quasiharmonic" approach, which is based on the quantum mechanical analysis of the inverse covariance matrix (Case 1994). Other than Hsu et al., they directly subtracted the absolute entropies of the two free proteins from the absolute entropy of the complexed system. They compared this method to normal mode calculations and noted that (1) absolute entropies did not converge over the time of the simulation, (2) their values were sensitive to the scheme used for the superpositioning of snapshots, and (3) the calculated overall loss of conformational entropy exceeded the harmonic estimate by 780 $\mathrm{cal\,mol^{-1}K^{-1}}$ (a free energy difference of 234 $\mathrm{kcal\,mol^{-1}}$) and was clearly unreasonable. Gohlke and Case diagnosed insufficient sampling and questioned the general applicability of the quasiharmonic approach.

3.3.4 The caveat of quasiharmonic analysis

As it turns out, the problem of Gohlke and Case is more related to the method than to the length of sampling. Using the same protocol, I derived absolute entropies S_{rec}, S_{lig} and S_{com} from the 10 x 1 ns ensembles of 7 free receptors, ligands and their complexes. Subtracting free from bound state should yield the change of conformational entropy induced by binding, that is

$$\Delta S_{conf} = S_{com} - (S_{rec} + S_{lig}) \; .$$

However, as in the example described by Gohlke and Case, the difference turned out far too negative for all 7 protein complexes (data not shown). Thus the method either systematically overestimates the entropy of the two free proteins or it sys-

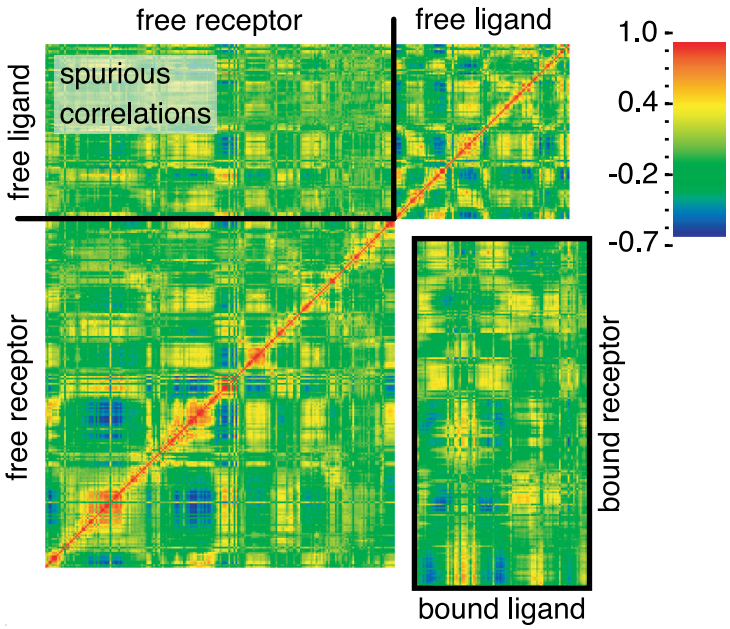

Figure 3.7: Spurious correlations between independent simulations. The trajectory of a free receptor is artificially combined with the trajectory of a free ligand. The correlation matrix of this large system of two independent molecules shows the expected intramolecular correlations in the lower left and upper right quadrant but also reveals unphysical correlations between the independent simulations (upper left quadrant). For comparison, the lower left quadrant shows real cross-correlations of the two molecules in the simulation of the protein complex. Data are taken from the last 500 ps of 3 single 1 ns simulations of c15. For space reasons, the plot only considers the x-coordinates of every 4^{th} atom.

tematically underestimates the entropy of the complex. The systematic nature of the error escaped the notice of Gohlke and Case, who examined a single protein complex.

There was only one difference between the calculations on free and complexed state that consistently applied to all 7 cases: The entropy calculated from one large system (the complex) was always compared to the value derived from two smaller systems (receptor and ligand). In order to eliminate this factor, I combined the two independent trajectories of free receptor and free ligand side by side into the single trajectory of an artificial complex. The artificial construct represented the free state but was the exact counterpart of the bound trajectory, both in terms of

length (10 x 1 ns) and number of atoms (size of receptor plus ligand). In theory, the conformational entropy S_{fcom} determined from the covariance matrix of this fake complex must be exactly equal to the sum $S_{rec} + S_{lig}$ of the independently calculated values. After all, there was never any exchange of information between the receptor and ligand ensemble, which were simply put next to each other. Nevertheless, in practice, S_{fcom} always turns out much lower, $S_{fcom} \ll S_{rec} + S_{lig}$. The reason for this puzzling result are spurious correlations between the independent simulations of receptor and ligand that, in principle, defy physical laws.

Figure 3.7 shows the normalized covariance matrix of two artificially combined trajectories. The lower left quadrant is identical to the covariance matrix obtained from the simulation of the free receptor. The atoms of this molecule exhibit correlated (yellow to red) as well as anti-correlated (blue) fluctuations. Both kinds of correlations reduce the entropy S_{rec}. The same holds for the covariance matrix of the free ligand simulation, which ends up in the upper right quadrant of the combined matrix. Surprisingly, the matrix also reveals (impossible) correlations between the independent molecules (upper left quadrant). These correlations lower the value of S_{fcom}. Their position in the matrix varies if different starting points or trajectories are chosen. Therefore they average out to some extent, but not completely, if the covariance matrix is constructed from 10 trajectories instead of 1. They are also lowered by longer simulation times, albeit to a lesser degree. Interestingly, the unphysical correlations disappear completely if I shuffle the time order of one molecule in the fake complex, that means $S_{fcom\_shuff} = S_{rec} + S_{lig}$.

Spurious correlations are probably an artifact of characteristic oscillations that are common to any protein. Vibrations of peptide planes, amide groups, or helix segments, to name only a few, are frequent throughout every deterministic simulation. In order to give a correlation, two coordinates have to move (1) at the same frequency and (2) in phase. Criterion (2) should prevent correlations across independent simulations. Yet, if some frequencies are indeed highly common, atoms moving at this pace in one simulation will often find atoms in another simulation (even of a different protein) that not only move at the same frequency but, by chance, also in phase. Surprising is the extent of such random correlations and

their persistence even in nanosecond simulations.

The entropy calculations described further below were based on a covariance matrix constructed from the last 300 ps of 10 independent 1 ns simulations. In this setup, spurious correlations lower S_{fcom} by several 100 cal mol^{-1}K^{-1}, about 4% to 5% of the absolute value. Unfortunately, several 100 cal mol^{-1}K^{-1} will always make a huge difference to the calculation of *relative* entropies between free and bound state.

3.3.5 Calculation of conformational entropies

The systematic underestimate of conformational binding entropies ΔS_{conf} is caused by spurious fluctuations between receptor and ligand in the complex (bound) trajectory, which are not considered if S_{rec} and S_{lig} are calculated independently for the free state. The problem is solved by calculating also the free entropies of receptor and ligand in a fake complex, that is $\Delta S_{conf} = S_{com} - S_{fcom}$. The strategy implies that the amount of spurious correlations remains the same for free and bound trajectories. Fortunately, this assumption can be tested. Any correlations that occur between the separately fitted receptor and ligand from two independent bound trajectories are nonphysical. The entropy S_{com_shift} of this fake complex (with receptor and ligand split along the interface and shifted by 1 ns) can be compared to the value $S_{brec} + S_{blig}$ calculated separately for the bound receptor and ligand ensemble. $S_{brec} + S_{blig}$ cannot contain any spurious correlations between receptor and ligand. Spurious correlations thus perturb the bound state by $S^{com}_{spurious} = S_{com\_shift} - (S_{brec} + S_{blig})$. As described above, the spurious correlations in the fake complex of the free state can be directly quantified as $S^{fcom}_{spurious} = S_{fcom} - (S_{rec} + S_{lig})$.

Indeed, the amount of spurious correlations in free and bound state is very similar for all protein complexes. For control, the difference $\Delta S_{spurious}$ is given along with the 7 entropy estimates in table 3.3. In summary, it appears justified to compare the entropy of a fake complex, assembled from the two free simulations, with the entropy of the real complex. I thus estimate the overall change of

conformational entropy as

$$\Delta S_{conf} = S_{com} - S_{fcom} \, .$$

There remains the problem of convergence. 10 x 1 ns simulations presumably provide better sampling of conformational space than the single 10 or 12 ns simulations used previously. Nevertheless, more or longer simulations could always turn up an additional region in conformational space that severely influences the entropy of the system. Several of the 10 x 1 ns simulation sets contained singular 1 ns trajectories that (judged by the rmsd to the start or end structure) kept diverging, whereas the remaining 9 trajectories appeared equilibrated over the last 500 ps. I excluded such "outlier" trajectories with an automatic procedure. In line with previous observations (Gohlke and Case 2004; Hsu et al. 2004), the absolute entropy of free or bound state did not converge. It generally increased with the adding of further simulation data – be it additional frames of a constant time segment (inset figure 3.8) or frames that covered a longer time (inset figure 3.9). By contrast, the entropy *difference* between bound and free state showed sufficient convergence. This concerns the density of sampling, shown in figure 3.8, as well as the necessary time coverage, shown in figure 3.9. Nevertheless, I should caution that the convergence backward in time could under some circumstances be more easily achieved than convergence with time.

For each complex, I selected a time segment and sampling interval, that insured a converged difference between bound and free entropy. For most complexes I settled on 10 x 300 ps coverage (taken from the simulation end) and 0.2 ps sampling interval. The largest complex, c20, was only sampled every 0.3 ps. The entropy calculated for complex c17 did not converge, the value given below was obtained from the last 10 x 500 ps with 0.3 ps sampling interval.

The quantification of spurious correlations already required the calculation of 5 additional entropies (S_{rec}, S_{lig}, S_{brec}, S_{blig}, S_{com_shift}) that were not needed for determining the overall change of conformational entropy. One more calculation provides further valuable information. The separate superpositioning of receptor and ligand in the complex trajectory (giving S_{com_split}) destroys the residual

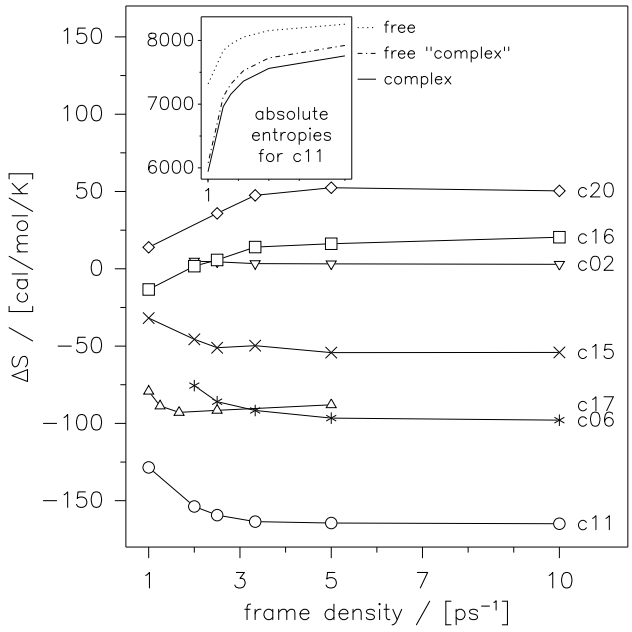

Figure 3.8: Dependency of conformational binding entropies on the sampling interval. Differences of bound and free conformational entropy were calculated for the last 10 x 200 ps (c17: 300 ps, c20: 100 ps) of simulations covering 10 x 1 ns but using different offsets between the snapshots (from 0.1 to 1 ps).

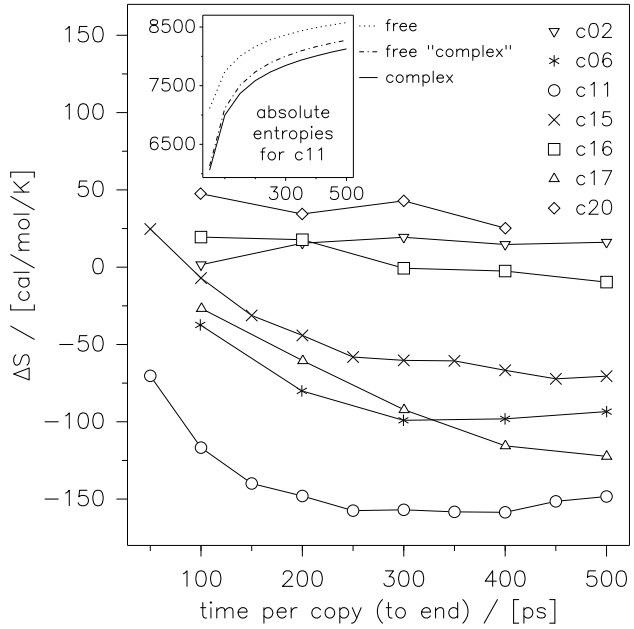

Figure 3.9: Convergence of binding entropies with simulation time. Absolute entropies of free or bound state do not converge (inset). However, the difference of entropies converges reasonably well for most protein complexes. Values were calculated as described for different time segments from the end of 10 x 1 ns simulations of free and bound state. For example, 100 ps translates to 10 segments covering the range 0.9-1 ns, 200 ps corresponds to 10 x 0.8-1 ns, and so on. Snapshots were taken every 0.2 or (c06, c17, c20) 0.3 ps.

Table 3.2: Protocols used for entropy calculations.

protocol	MD[1]	analyze[2]	fit[3]	pair[4]	order[5]
rec	free	rec	r		
lig	free	lig	l		
fcom	free	rec+lig	r ‖ l	2	intact
fcom_shuff	free	rec+lig	r ‖ l	2	shuffled
brec	com	rec	r		
blig	com	lig	l		
com	com	rec+lig	r + l	1	intact
com_split	com	rec+lig	r ‖ l	1	intact
com_shift	com	rec+lig	r ‖ l	2	intact
com_shuff	com	rec+lig	r ‖ l	1	shuffled

[1]source simulation: free receptor and/or ligand (free), complex (com); [2]consider receptor (rec) and/or ligand (lig); [3]fit trajectories on reference rec (r) and lig (l), seperately (r ‖ l), or as single molecule (r+l); [4]rec and lig taken from same (1) or two independent (2) simulations; [5]time order of ligand coordinates;

motion of the two molecules. It therefore reveals the entropy content of these intermolecular fluctuations (denoted rec⟆lig in table). The remaining correlations across the binding interface are uncovered if the complex trajectory is moreover divided into receptor and ligand and reassembled from two independent bound trajectories (S_{com_shift} described above). The difference between S_{com_split} and S_{com_shift} (denoted rec×lig in table) yields the negative entropy contribution from correlated motions between receptor and ligand. Table 3.2 summarizes the different entropy values calculated for each complex.

The analysis of 10 independent 1 ns trajectories instead of one 10 ns simulation allows for a strict error estimate. I repeated all entropy calculations 5 times, excluding different tiers of three trajectories from the analysis.

3.3.6 Conformational entropy of binding

Table 3.3 provides the total conformational entropy of binding, ΔS_{conf}, calculated for 7 protein complexes. The values cover a wide range from stark entropy loss

Table 3.3: Conformational binding entropy and its decomposition calculated from quasi-harmonic analysis. All values are for bound - free state in cal/mol/K.

	all heavy atoms						backbone[7]
	rec[1]	lig[2]	rec⌇lig[3]	rec×lig[4]	$\Delta^5_{spurious}$	ΔS_{conf} [6]	ΔS_{conf} [6]
c02	20 ±37	-46 ±18	46 ±1	-11 ±0	2.4±3	19 ±44	19 ±22
c06	-196 ±43	31 ±20	55 ±1	-8.2±1	4.3±2	-101 ±34	12 ±12
c11	-115 ±35	-123 ±33	55 ±3	-8.9±0	3.6±2	-157 ±41	-27 ±14
c15	-55 ±24	-103 ±22	70 ±2	-6.6±0	3.0±1	-60 ±25	12 ±8
c16	75 ±15	-144 ±17	45 ±1	-13 ±1	3.0±2	-0.7±13	40 ±5
c17	-61 ±40	-149 ±23	59 ±3	-13 ±1	5.5±5	-122 ±40	7.0±14
c20	-149 ±18	104 ±16	60 ±2	-14 ±1	4.4±7	43 ±23	17 ±9

[1]receptor only; [2]ligand only; [3]entropy gain from rigid body motions of receptor against ligand; [4]entropy loss from motions correlated across the binding interface; [5]difference between spurious correlations in free and bound state (see text); [6]total change of vibrational entropy; [7]only considering carbonyl C and O

(-157 cal mol^{-1}K^{-1}) to substantial gain (43 cal mol^{-1}K^{-1}). The two smallest complexes, c11 and c15 exhibit a large entropy loss, as does the antibody antigen system (c06). The largest assembly, c20, yields the highest gain of conformational entropy. Two other comparatively large systems, c02 and c16, give a moderately positive or unchanged entropy. Another large complex, c17, seems to assemble at a high cost of conformational entropy. However, the latter value has to be treated with caution. The entropy calculations of c17 did not converge and the receptor ensemble constituted an outlier in the analysis of flexibilities (see 3.3.1 on page 65).

Table 3.3 also gives the (pseudo) entropy difference between bound and free state of the isolated receptor and ligand proteins. The individual partners can both gain or loose conformational entropy upon binding. Nevertheless, a loss of conformational entropy seems to be more common than perhaps expected from the analysis of conformational diversity (section 3.3.2). On the other hand, the selection of 7 complexes was, from the beginning, biased toward reduced flexibility of the bound state (compare figures 3.6 and 3.5). There appears to be a trend toward entropy compensation between receptor and ligand. The three complexes

with positive or unchanged ΔS_{conf} always combine the large entropy gain of one protein with a sizable loss of entropy on the other side.

Nevertheless, the simple summation of receptor and ligand values does not yield the overall conformational entropy of binding. The simplistic approach ignores an important component: Even in the complex, the two molecules move with respect to each other. This intermolecular motion translates into a substantial entropy gain (rec$\wr$lig in table 3.3) of, generally, about 50 cal mol^{-1}K^{-1} or more. The value is in remarkable agreement with a "guess" (Janin 1995) made by Finkelstein and Janin (1989) over 15 years ago. The simple summation of receptor and ligand entropies also ignores the negative contribution arising from correlations of motion across the binding interface (rec$\times$lig in table 3.3). Compared to other contributions, there is only moderate crosstalk between receptor and ligand fluctuations. The smallest complex "looses" 6.6 cal mol^{-1}K^{-1}, the largest 14 cal mol^{-1}K^{-1} to such correlations.

The large error margin of ΔS_{conf} in table 3.3 testifies to the deficiencies of conformational sampling. Interpreted on a positive note, the margin may thoroughly capture this deficiency because it stems from the comparison of truly independent simulations. Moreover, the error estimate was calculated with a 30% smaller data set than the reference value, and should thus be on the conservative side. Errors of about $\pm$30 cal mol^{-1}K^{-1} would introduce an uncertainty of 9 kcal mol^{-1} K^{-1} to calculations of binding free energies. However, the total stability of protein complexes typically ranges only from 6 to 15 kcal mol^{-1} K^{-1}. The accuracy of the method must thus be substantially improved before it can serve practical purposes.

Despite many uncertainties, my calculations demonstrate that the change of conformational entropy should have a considerable influence on the overall stability of protein complexes. It is too early, perhaps even fundamentally impossible, to make general statements as to the sign of this contribution. Protein association can not only deplete but also boost conformational entropy. Especially larger complexes seem sometimes able to compensate for the loss of diversity occurring in the binding region.

It would be helpful to cross-check the predicted change of conformational

Table 3.4: Computational estimates and (where available) experimental measurements of overall binding entropies. All values are for bound - free state in cal/mol/K.

	ΔS_{conf} [1]	$\Delta S_{t,r}$ [2]	ΔS_{solv} [3]	ΔS_{total} [4]	ΔS_{exp} [5]	Δres [6]	ref [7]
c02	19 ±44	-100	334	253 ±44		0	
c06	-101 ±34	-105	135	-71 ±34	(-34)[8]	0	[1]
c11	-157 ±41	-101	242	-12 ±41	-1	2	[2]
c15	-60 ±25	-98	207	49 ±25	20	2	[3]
c16	-0.7 ±13	-103	422	318 ±13		-1	
c17	-122 ±40	-108	587	357 ±40		-4	
c20	43 ±23	-109	556	490 ±23		12	

[1]total change in conformational entropy; [2]loss of rotational and translational entropy; [3]change in solvent entropy estimated from buried accessible surface; [4]total entropy change calculated (1+2+3); [5]measured entropy change where available; [6]number of residues that are disorded in the free and ordered in the bound (-, vice versa); [7]literature for 5: [1] (Bhat et al. 1994; Schwarz et al. 1995; Sundberg et al. 2000), [2] (Frisch et al. 1997), [3] (Arold et al. 1998); [8]measured for a related molecule

entropies with experimental values. Direct measurements of conformational entropy are often attempted by determining order parameters of the peptide bond plane from NMR relaxation experiments. For comparison, table 3.3 also provides the change of conformational entropy that I calculate if only the fluctuations of backbone carbonyl carbon and oxygen are considered. Changes of overall and backbone entropies can differ substantially. They are correlated with R=0.8 and the backbone calculation is strongly biased toward gains of entropy. Backbone atoms seldom make direct contacts with the other binding partner and are hence less likely to loose mobility upon binding. They nevertheless benefit just like side chain atoms from flexibility gains outside the contact interface. Measurements that are only based on backbone atoms have thus to be interpreted with care. Unfortunately, there is no dependable experimental reference quantity for conformational entropy changes.

3.3.7 The overall entropy cost (or gain) of binding

Overall entropy changes due to binding can be reliably measured by calorimetry. Corresponding data are available for the binding of Barnase to Barstar (c11) as well as HIV-1 $\text{Nef}_{\Delta 1-57}$ to the SH3 domain of Fyn (c15). Data have also been published for the interaction between the mouse antibody fragment FvD1.3 and hen egg white lysozyme but not for the variant studied here (c06) which involves an artificial hybrid of FvD1.3 and human antibody segments (Holmes et al. 1998). The comparison of experimental and theoretical entropies of c11 and c15 is complicated by unresolved residues in the structures of both systems. 18 terminal and 30 non-terminal residues of the free as well as 15 terminal and 31 non-terminal residues of the bound $\text{Nef}_{\Delta 1-57}$ are disordered and hence not present in my simulations. The putative disorder-order transition of effectively two residues of $\text{Nef}_{\Delta 1-57}$ and two terminal residues within the Barnase / Barstar complex is also not reflected in my entropy calculations. Salt concentrations, different protonation states and other details may introduce further inaccuracies.

In contrast to my calculations, the experimental values include the entropic contribution from the solvent. The change of solvent free energy is commonly estimated from the accessible surface area ΔASA buried upon folding or binding (Brady and Sharp 1997): $\Delta G_{solvent} = \gamma \Delta ASA$. It is assumed to be largely of entropic nature (Sharp et al. 1991). I here used $\gamma = 47$ cal $\text{mol}^{-1}\text{Å}^{-2}$ (Sharp et al. 1991; Noskov and Lim 2001). Table 3.4 combines the changes of vibrational, rotational and translational entropies given in table 3.3 with the estimated gain of solvent entropy and compares this overall value with the three available experimental binding entropies. For Barnase / Barstar (c11) the experimental value falls well within the (broad) 68% confidence interval of the calculated entropy. For c06 the computed value is about one standard deviation below- and for c15 1.2 standard deviations above the experimental observation. However, compared to mouse FvD1.3, the humanized antibody requires additional conformational adjustments to bind its target (Holmes et al. 1998) and the entropy loss of c06 may be indeed larger than measured for the complex of mouse FvD1.3. Moreover, to ignore the putative ordering of two residues upon binding may have overestimated

the values computed for c11 and c15. The same applies almost certainly to c23 (lacking a experimental reference).

Also the estimate of solvation free energies may be prone to significant error. The proportionality constant γ is derived from the distribution of model compounds between polar and unpolar solvents (Brady and Sharp 1997). Early works put it at lower values and other studies considered only the change of unpolar surface. Jackson and Sternberg (1995) argued that the molecular surface is a better measure than the accessible surface. More recently, Kyte (2003) showed that solvation free energies of small molecules correlate better with the number of exposed hydrogen-carbon bonds rather than with a surface based measure. He derived a value of 20 KJ mol^{-1}C-H^{-1} but did not elaborate on how to divide exposed from buried hydrogen-carbon bonds in larger molecules. Considering the exposed surface of a CH_2 moiety, his method translates again to about 50 cal $mol^{-1}Å^{-2}$ but, depending on the surface composition, it may give different overall estimates.

In spite of many uncertainties, my calculations are at least compatible with the experimentally observed binding entropies. It remains to be seen whether such agreement of experiment and theory (at very low level of precision) also holds for other complexes.

3.4 Recognition between structure ensembles

3.4.1 Definition of ensembles

In the following, we try to incorporate the additional dimensions of receptor and ligand variability into the picture of the protein-protein recognition process. This recognition starts from the unbound components and we therefore once more concentrate on the conformational ensembles of the free receptor and the free ligand.

As explained in section 3.2.3, we had probed the motion of all 33 unbound proteins with short (10 x 50 ps) molecular dynamics simulations (see section 3.2.3). Unfortunately, large-scale correlated motions usually escape the sampling of MD simulations (Balsera et al. 1996). We therefore calculated a second set of 33 en-

sembles with identical protocol except of a weak restraint alleviating this problem. Large-scale correlated motions typically occur along small gradients in the energy landscape. They are hence slow but, on the other hand, can be boosted by small interventions. As described previously (Abseher and Nilges 2000), the restraint acts on the ensemble of 10 concurrent trajectories as a whole and increases the variability along the major principal components of motion. The computational cost of this principal component restrained simulation (PCR-MD) is similar to the classic approach above but the ensemble is considerably more diverse.

We performed c-means fuzzy clustering for each of the two structure ensembles and selected 2 x 10 representative conformations for combinatoric rigid body docking. A representative example of these discretized structure ensembles from the unrestrained (MD) and the restrained (PCR-MD) simulation is shown in figure 3.10. The snapshots capture considerable variation. Table 3.5 lists the average (rms) deviation between the members of each docking ensemble and their distance to the free and the bound structure.

3.4.2 Ensemble docking

We tried to mimic the recognition between two flexible molecules by a combinatoric docking of all snapshots from the receptor ensemble against all snapshots from the ligand ensemble. Each of the docking ensembles was supplemented with the free (experimental) structure. Using the docking program HEX (Ritchie and Kemp 2000), we performed 121 rigid body dockings for each complex and MD strategy. HEX represents receptor and ligand by a soft 3D surface skin model and calculates the volume of water that is expelled from the protein surfaces as they come together. In addition there is a penalty for steric overlap. Both terms are combined in a pseudo energy that depends solely on the atomic and water probe radii and is interpreted as an approximation of the desolvation and van der Waals component of the free energy of association. We did not employ any additional (e.g. electrostatic) potentials and dealt therefore only with the contribution of short range, geometry dependent, effects to the interaction free energy. HEX performs a systematic search over all 6 rigid body degrees of freedom and ranks

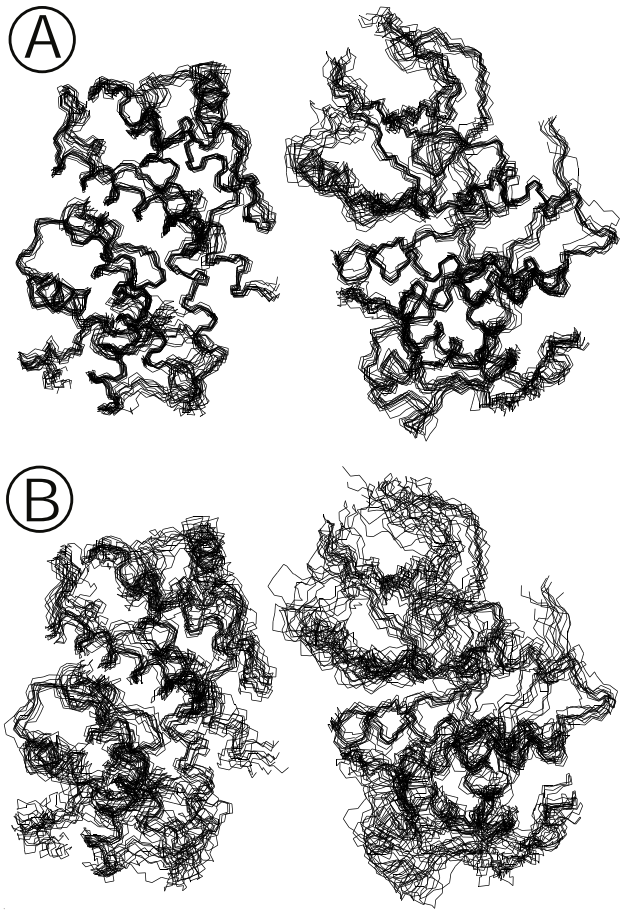

Figure 3.10: Receptor and ligand ensembles used for the docking of c20. (A) The 10 receptor (right) and 10 ligand (left) snapshots selected from the unrestrained simulations. (B) The 10 snapshots from the principle component restrained simulations (PCR-MD) cover a wider range of conformations. The receptor and ligand snapshots have been oriented as in the native complex, but are separated horizontally. Side chains have been omitted for clarity.

Table 3.5: Average rmsd of structure ensembles.

ID	PDB code	MD		PCR-MD		interface RMSD to bound[c]	
		pairwise[a]	to free[b]	pairwise[a]	to free[b]	MD	PCR-MD
c01	1BRA	1.3±0.09	1.2±0.16	2.9±0.52	2.1±0.56	1.6±0.12	2.2±0.48
	1AAP	1.5±0.21	1.4±0.28	2.1±0.35	1.7±0.35	1.6±0.26	1.6±0.17
c02	2CGA	1.3±0.08	1.2±0.16	2.6±0.41	1.9±0.39	3.0±0.08	3.4±0.30
	1HPT	1.6±0.19	1.5±0.25	2.3±0.37	1.8±0.36	3.0±0.19	3.1±0.19
c03	2PKA	1.4±0.09	1.5±0.18	2.8±0.65	2.1±0.58	2.2±0.25	2.5±0.50
	5PTI	1.5±0.22	1.4±0.29	1.9±0.25	1.6±0.33	1.6±0.26	1.5±0.21
c04	1SUP	1.3±0.10	1.2±0.18	2.8±0.48	2.1±0.48	1.5±0.17	2.4±0.45
	3SSI	1.4±0.11	1.4±0.24	2.1±0.28	1.7±0.28	1.8±0.24	1.9±0.30
c05	1FGN	1.7±0.14	1.7±0.30	2.9±0.47	2.4±0.53	1.7±0.23	1.9±0.30
	1BOY	1.7±0.12	1.6±0.28	2.6±0.40	2.1±0.48	1.7±0.18	2.0±0.32
c06	1BVL	1.5±0.09	1.4±0.17	2.7±0.46	2.1±0.38	1.8±0.19	2.2±0.37
	3LZT	1.2±0.10	1.1±0.21	2.5±0.47	1.9±0.53	2.4±0.22	2.7±0.28
c08	1DQQ	1.5±0.13	1.6±0.29	2.6±0.40	2.1±0.47	1.5±0.15	1.6±0.18
	3LZT	1.2±0.10	1.1±0.21	2.3±0.41	1.8±0.38	1.9±0.14	2.3±0.28
c11	1A2P	1.3±0.10	1.2±0.19	2.2±0.38	1.8±0.55	1.7±0.21	2.3±0.62
	1A19	1.5±0.11	1.4±0.15	2.4±0.40	1.9±0.34	1.5±0.15	1.8±0.30
c13	2BNH	1.5±0.10	1.6±0.24	2.8±0.41	2.2±0.46	2.7±0.36	2.9±0.61
	7RSA	1.5±0.13	1.4±0.26	2.3±0.38	1.9±0.46	1.9±0.22	2.4±0.43
c14	1VXR	1.4±0.07	1.4±0.22	3.1±0.58	2.3±0.56	2.1±0.24	2.5±0.50
	1FSC	1.5±0.17	1.4±0.23	2.2±0.36	1.8±0.43	2.0±0.29	2.3±0.38
c15	1AVV	1.6±0.17	1.5±0.23	2.6±0.42	2.0±0.35	1.5±0.18	1.8±0.24
	1SHF	1.5±0.16	1.4±0.22	1.8±0.19	1.6±0.23	2.0±0.21	2.1±0.19
c16	1AKZ	1.3±0.07	1.2±0.17	2.0±0.28	1.6±0.30	1.7±0.18	1.8±0.27
	1UGI	1.6±0.18	1.4±0.25	2.4±0.40	1.8±0.41	1.8±0.14	2.1±0.31
c17	1WER	1.5±0.10	1.5±0.22	2.3±0.36	1.8±0.37	1.7±0.08	2.0±0.30
	5P21	1.4±0.08	1.3±0.21	2.4±0.35	1.9±0.41	2.4±0.28	2.8±0.47
c19	1PIF	1.4±0.06	1.3±0.22	3.1±0.65	2.2±0.61	2.0±0.29	2.6±0.51
	2AIT	1.6±0.15	1.6±0.20	2.4±0.33	2.1±0.34	2.0±0.16	2.1±0.25
c20	1HCL	1.6±0.12	1.5±0.22	2.7±0.40	2.0±0.38	7.9±0.19	8.0±0.38
	1VIN	1.4±0.07	1.4±0.20	2.4±0.33	1.8±0.36	1.7±0.14	1.9±0.28
c21	1B39	1.6±0.11	1.4±0.18	2.3±0.39	1.9±0.40	5.8±0.16	5.8±0.35
	1FPZ	1.6±0.10	1.6±0.23	2.4±0.35	2.0±0.35	2.4±0.21	2.6±0.31
c22	1TBG	1.6±0.14	1.5±0.22	3.4±0.62	2.6±0.58	1.6±0.19	2.2±0.44
	1TAG	1.5±0.09	1.4±0.26	2.4±0.33	1.9±0.41	6.4±0.10	6.4±0.20

[a] Average pairwise heavy atom rmsd in (with standard deviation) between the 10 simulation snapshots.
[b] Average heavy atom rmsd in (with standard deviation) of the 10 simulation snapshots to the free structure.
[c] Average heavy atom rmsd (with standard deviation) of interface residues between the 10 simulation snapshots and the bound structure.

in the order of 109 trial orientations by this interaction energy.

From each of the 121 HEX dockings we analyzed the 512 top ranking solutions provided by default. Since we did not apply any clustering and there was no random element in the search, the amount and quality of near-native orientations within the set of top-ranking solutions effectively depended on: (1) how well the two protein conformations matched each other geometrically near the native orientation, (2) how tolerant this steric match was to deviations from the optimum orientation, and (3) how many non-native alternative orientations with comparable geometric match existed and competed with the correct arrangement.

3.4.3 Measuring the quality of docking solutions

We analyzed and compared 2,106,368 solutions from 4114 rigid body docking calculations between 693 conformations of 33 different proteins (c06 and c08 share a ligand). To this end we needed a single metric for the quality of a given solution, that is to which extent it resembles the native arrangement of receptor and ligand in the complex. Rmsd-based measures are inappropriate for our purposes because they depend on the size and shape of the binding interface and, furthermore, would also be influenced by the conformational variations in our receptor and ligand ensembles. Criteria based on residue-residue contacts (Mendez et al. 2003) suffer from ambiguity introduced by bulky side chains in the interface. We therefore used a measure based on atom contacts. We define a fraction of native atom contacts (fnac) as the number of pairs of non-hydrogen receptor and ligand atoms that are within a 10 Å distance both in the native and the predicted orientation, divided by the total number of such pairs in the native complex.

3.4.4 Complementarity across ensembles

Discrimination by shape complementarity alone is usually sufficient to predict the native arrangement of the bound receptor and ligand. In figure 3.11A the docking of the bound structures from c19 (glycosyltransferase / tendamistat) is shown as a representative example. The free structures, on the other hand, are generally

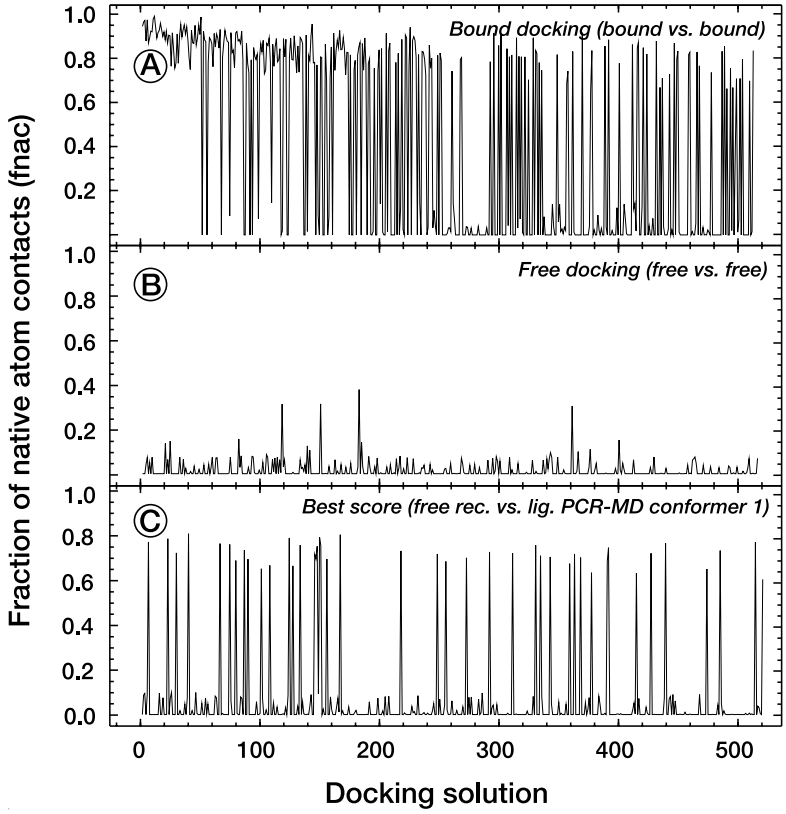

Figure 3.11: Selected docking results for c19. Each panel shows the result of a single shape-driven rigid body docking experiment. The similarity to the true complex is measured for the 512 predictions that rank highest in surface complementarity. Data are shown for (A) bound docking, (B) free docking and (C) the highest scoring of the ensemble dockings (see table 3.6).

much less complementary. For example, the majority of top-ranking solutions from the docking of free glycosyltransferase and tendamistat (figure 3.11B) reproduce no, or only few, native contacts. However, figure 3.11C shows the fnac (quality) of top-ranking solutions from the docking of the same free receptor structure against one of the alternative inhibitor conformations from the PCR-MD simulation. Clearly, this combination of structures had a better geometric fit in near native orientations. In figure 3.12A we show the amount and quality of near native solutions for all cross-dockings between the simulation-derived ensembles of the two proteins. Several conformer combinations performed better than the docking of the two experimental structures, both in terms of quantity (indicated by the size of the circle) and quality (indicated by the color). The gain was yet even more pronounced for the cross-docking of the ensemble that had been calculated with the PCR-MD technique (figure 3.12B).

As a second example we present similar results for the complex between CDK2 and cyclin A (c20). This complex is one of the difficult docking test cases as the receptor undergoes large structural changes moving from the free to the bound state (C_α displacements of up to 20 Å). All 512 solutions from the docking of the two experimental structures stayed below a fnac of 10%. Nevertheless, as shown in figure 3.12B and C there were many combinations of MD or PCR-MD snapshots that yielded better solutions with fnac values up to 30%.

The results of all 17 test complexes are provided in supplemental figure S3 and summarized in table 3.6 and figure 3.6. We selected 2 dockings each from the cross-docking of MD and of PCR-MD ensembles: The one that generated the single highest fnac within the 512 top-ranking orientations and the one with the best compromise between quantity and quality of near-native solutions. We quantify this "compromise" docking performance with the sum of squared fnac values above 10%, i.e. a simple score strongly biased toward high fnac ranges.

The cross-docking of ensemble snapshots always found more and, in all but one case, also better near native solutions than the docking of the free conformations alone. There were usually several combinations of simulation snapshots, or snapshot and free structure, with better complementarity near the native orienta-

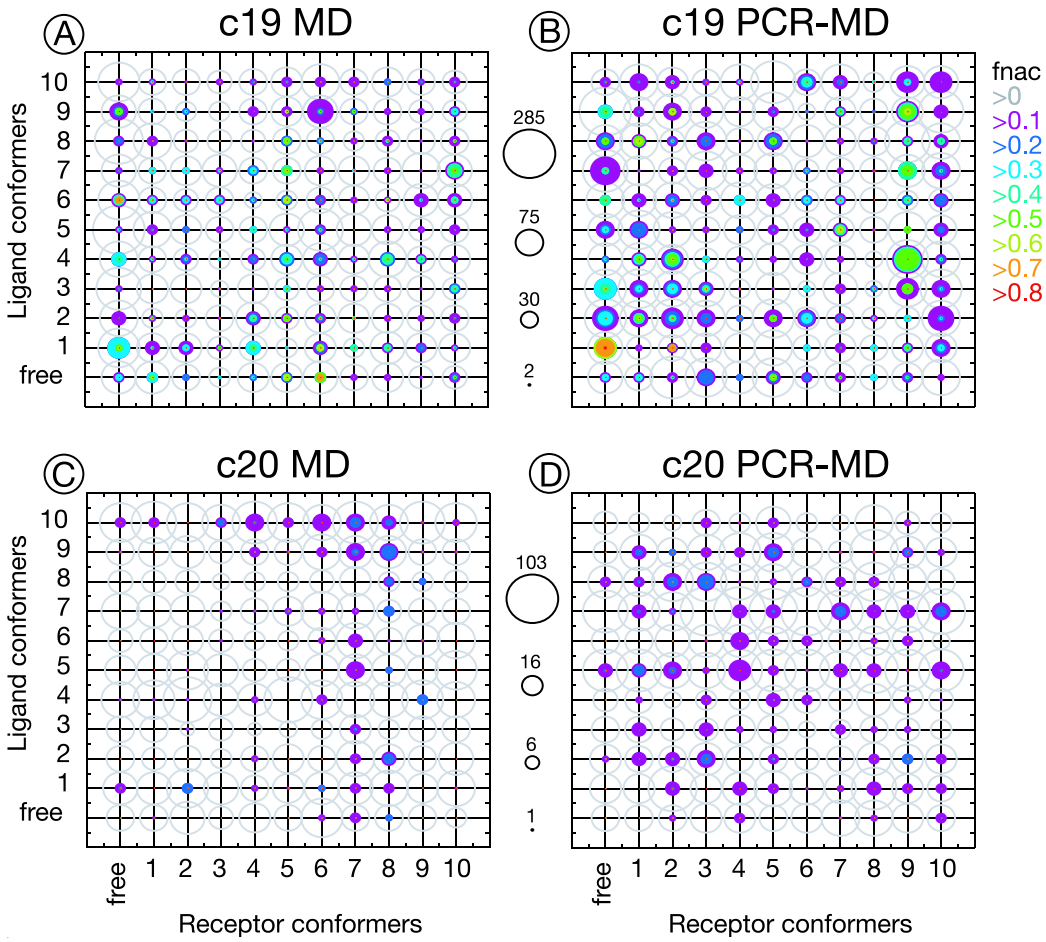

Figure 3.12: Quantity and quality of near-native solutions in 4 selected ensemble dockings. The cross-docking of 11 receptor and 11 ligand conformations generates 121 sets of 512 docking solutions. The amount and quality of near native solutions among each set is shown for the ensemble dockings of c19 (A and B) and c20 (C and D). The area of each contour is proportional to the number of solutions (see the separate size legends). The color of a contour indicate solutions above a certain fnac-value (see the color legend). Several conformer-combinations perform better than the traditional docking of the free structures.

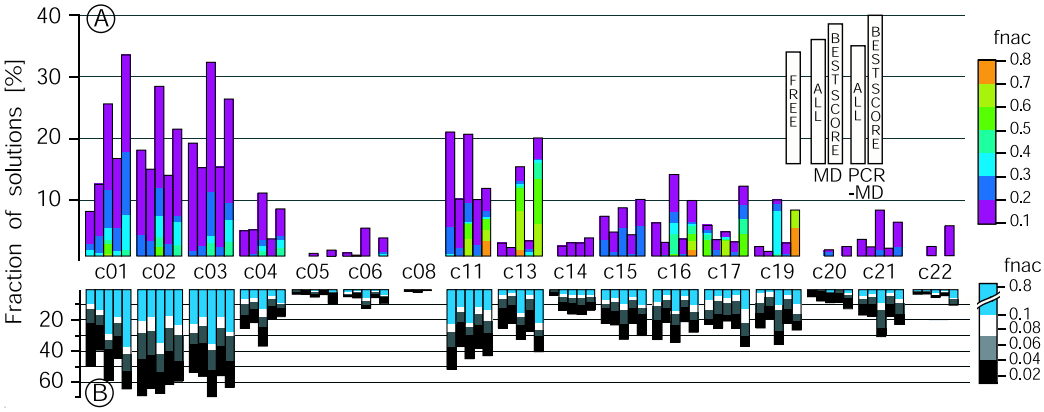

Figure 3.13: Quantity and quality of near-native solutions in all test cases. The amount of solutions above a certain quality (fnac) level (see color legend) is given for selected docking runs of all 17 test complexes. Data for each complex are presented using groups of five bars. The first bar describes the free docking (512 orientations), the second and fourth bar show the data for all cross dockings (11 x 11 x 512 orientations). Bars three and five show the data for the best performing conformer-combination (512 orientations) from the MD and PCR-MD ensemble, respectively (see table 3.5). The upper plot (A) depicts solutions with fnac above 10%, while the lower plot (B) uses a 1% fnac threshold.

tion. Moreover, we can assume that even better fits remained hidden due to the fact that our docking ensembles were artificially sparse. The insufficient shape complementarity between many of the free receptor and ligand pairs could be an artifact of the rigid body or rigid backbone simplification.

3.4.5 Specificity of docking success

For every complex, we generated 10 random orientations that were distinct both from each other and the native (no contact overlap). We re-analyzed all docking solutions using these random orientations as reference. This allowed us to quantify the probability that the score of the free docking and the best score from the ensemble docking did not occur at random (table 3.6). All of the best performing conformer pairs reproduced the native complex better than the docking of the free experimental structures. In 9 out of 17 cases, the profound enrichment of high quality solutions from the docking of selected conformer pairs is also specific to

Table 3.6: Docking results for all test complexes.

ID	FREE				HIGHEST FNAC						BEST SCORE						average
		interf. rms[a]				conformer		interf. rms[a]				conformer		interf. rms[a]			
	fnac[b]	rec	lig	score[c,e]		fnac[b]	rec	lig	rec	lig	score[e]	fnac[b]	rec	lig	rec	lig	score[c,e]	score[d,e]
c01	0.82	1.4	1.5	3.9 *(88+)*	MD	0.88	6	8	1.6	1.5	3.5	0.68	Free	6	1.4	2.2	10.2 *(85+)*	2.9 *(74+)*
					PCR-MD	0.86	Free	2	1.4	1.4	5.3	0.56	3	4	2.5	1.6	10.9 *(78+)*	3.8 *(84+)*
c02	0.35	2.9	2.8	2.9 *(97+)*	MD	0.66	9	2	3.1	2.8	3.3	0.61	6	3	3.1	3.1	10.4 *(96+)*	2.6 *(98+)*
					PCR-MD	0.63	Free	8	2.9	2.8	2.7	0.48	1	6	3.7	3.0	7.0 *(86+)*	2.3 *(97+)*
c03	0.75	1.3	1.4	4.2 *(83+)*	MD	0.75	Free	Free	1.3	1.4	4.2	0.53	Free	3	1.3	1.6	8.2 *(81+)*	2.5 *(80+)*
					PCR-MD	0.75	Free	Free	1.3	1.4	4.2	0.54	Free	4	1.3	1.7	8.5 *(68+)*	2.2 *(79+)*
c04	0.22	0.9	1.3	0.5 *(58+)*	MD	0.77	Free	1	0.9	1.5	0.9	0.70	6	4	1.3	1.5	4.3 *(71+)*	1.1 *(78+)*
					PCR-MD	0.74	10	5	1.3	1.5	0.9	0.50	9	8	2.4	1.9	3.9 *(57+)*	0.6 *(54+)*
c05	0.09	1.1	1.2	—	MD	0.20	6	2	1.8	1.6	0.2	0.20	6	2	1.8	1.6	0.2 *(96-)*	0.0 *(81-)*
					PCR-MD	0.23	3	1	1.7	2.3	0.1	0.18	7	5	2.3	2.5	0.3 *(96-)*	0.0 *(92-)*
c06	0.16	1.4	2.3	0.1 *(3-)*	MD	0.73	8	6	2.0	2.4	0.8	0.35	2	1	2.1	2.3	0.8 *(35-)*	0.1 *(26-)*
					PCR-MD	0.69	2	10	1.7	2.5	1.0	0.40	3	1	2.1	2.3	1.1 *(90-)*	0.1 *(38-)*
c08	0.09	1.2	1.7	—	MD	0.39	6	4	1.5	1.9	0.2	0.39	6	4	1.5	1.9	0.2 *(2-)*	0.0 *(1-)*
					PCR-MD	0.26	Free	6	1.2	2.0	0.2	0.26	Free	6	1.2	2.0	0.2 *(3-)*	0.0 *(1-)*
c11	0.75	1.0	1.0	4.8 *(81+)*	MD	0.81	Free	10	1.0	1.5	14.9	0.81	Free	10	1.0	1.5	14.9 *(88+)*	2.1 *(71+)*
					PCR-MD	0.82	Free	3	1.0	1.9	7.6	0.81	10	9	2.6	1.1	17.2 *(90+)*	2.7 *(79+)*
c13	0.76	1.8	1.2	0.9 *(64+)*	MD	0.83	3	4	2.0	1.8	7.0	0.78	10	3	2.6	1.8	26.1 *(99+)*	1.4 *(78+)*
					PCR-MD	0.83	Free	6	1.8	1.5	3.2	0.67	5	4	2.6	2.3	8.0 *(69+)*	0.7 *(58+)*
c14	0.15	1.4	1.7	0.1 *(2-)*	MD	0.84	8	8	2.0	1.8	2.2	0.84	8	8	2.0	1.8	2.2 *(19-)*	0.3 *(27-)*
					PCR-MD	0.69	6	6	2.3	2.2	2.4	0.69	6	6	2.3	2.2	2.4 *(15-)*	0.4 *(25-)*
c15	0.27	1.1	1.7	1.5 *(45+)*	MD	0.54	4	7	1.4	2.1	0.7	0.31	4	5	1.4	1.8	2.2 *(43-)*	0.9 *(5-)*
					PCR-MD	0.70	2	7	1.7	2.1	1.2	0.54	7	Free	2.2	1.7	2.6 *(30-)*	0.8 *(10-)*
c16	0.73	1.2	1.7	1.8 *(80+)*	MD	0.82	Free	7	1.2	1.6	3.0	0.59	5	10	1.5	1.8	7.5 *(83+)*	1.1 *(64+)*
					PCR-MD	0.85	4	1	1.8	1.5	12.9	0.85	4	1	1.8	1.5	12.9 *(91+)*	1.4 *(70+)*
c17	0.70	1.5	1.7	6.9 *(100+)*	MD	0.80	3	Free	1.8	1.7	2.8	0.74	10	Free	1.6	1.7	8.3 *(99+)*	1.5 *(97+)*
					PCR-MD	0.77	3	Free	1.5	1.7	3.1	0.65	Free	2	1.5	2.0	11.0 *(100+)*	1.3 *(99+)*
c19	0.39	1.2	1.7	0.6 *(49+)*	MD	0.77	Free	6	1.2	1.8	3.4	0.62	Free	1	1.2	1.8	6.8 *(72+)*	0.7 *(57+)*
					PCR-MD	0.81	Free	1	1.2	1.8	22.4	0.81	Free	1	1.2	1.8	22.4 *(92+)*	1.4 *(74+)*
c20	0.09	7.8	1.4	—	MD	0.29	8	7	7.8	1.6	0.3	0.27	8	9	7.8	1.6	0.5 *(38-)*	0.1 *(37-)*
					PCR-MD	0.30	5	9	8.1	2.1	0.5	0.30	5	9	8.1	2.1	0.5 *(46-)*	0.1 *(19-)*
c21	0.24	5.9	1.9	0.5 *(75+)*	MD	0.36	8	7	5.7	2.3	0.9	0.32	8	8	5.7	2.6	1.2 *(5-)*	0.3 *(45-)*
					PCR-MD	0.42	Free	2	5.9	2.6	0.6	0.28	1	7	5.7	2.6	1.1 *(16-)*	0.3 *(41-)*
c22	0.15	1.0	6.2	0.0 *(19-)*	MD	0.22	5	5	1.9	6.6	0.2	0.22	5	5	1.9	6.6	0.2 *(1-)*	0.0 *(16-)*
					PCR-MD	0.26	6	4	2.4	6.4	0.3	0.18	6	10	2.4	6.7	0.6 *(1-)*	0.1 *(23-)*

[a] heavy atom interface rmsd to the bound structure.
[b] fraction of native atom contacts (highest of the docking run).
[c] docking performance score, defined as the sum of squared fnac-values above 0.1, quantifies the amount and quality of near-native solutions captured by the docking run.
[d] average docking score of all 121 dockings.
[e] probability (in percent) that the value is an outlier above (+) or below (-) random expectation. Probabilities for scores around or below random expectations are afflicted with higher error due to the asymmetric shape of the lognormal random distribution. Probabilities also depend on the number of scores considered. The best of 121 combinatorical dockings must accordingly perform much better than the single free docking in order to achieve the same significance (specificity).

the native orientation. In the remaining cases, the improvement is substantial but not significantly higher than what would be expected for a random orientation. We have indications that more specific results can be achieved for some of the 8 latter complexes if the HEX energy function is extended with an electrostatic term.

It should be noted that the consideration of 512 solutions each from 121 docking runs combined with the soft and simplistic energy function provoke a high level of "noise", i.e. similarities to a random orientation. The evaluation of fewer solutions with more detailed energy functions would most likely improve the discrimination. However, the technical (and challenging) problem of scoring docking solutions is not subject of this work.

3.4.6 Recognition conformations

Our simulations cover a time window that, at least, resembles but probably exceeds the estimated duration of a micro-collision. Already the use of multiple trajectories is expected to increase sampling by a factor of 2 (Caves et al. 1998). The fast equilibration, the method of solvation and, especially, the introduction of principle component restraints further enhance diversity (Abseher and Nilges 2000). We did not find a global transition from free to bound interface conformation in any of our 2 x 33 ensembles. There was nevertheless notable variation in the structure ensembles and some conformers were necessarily closer to the bound than others (compare table 3.5). Binding could be promoted by such shifts toward the bound state (Kumar et al. 2000). In figure 3.14A we relate the distance from the bound state of a given pair of conformers and its performance in docking. There is no obvious correlation between similarity to the interface of the bound structure and docking performance. This picture remained the same when we expressed the distance between structures as Contact Area Difference (Abagyan and Totrov 1997) (data not shown) and is therefore not an artifact of the rmsd measure.

In figure 3.14B and C we focus only on those pairs of conformations that yielded the best docking result (score) for each complex. As apparent from table 3.6, the experimental structure was over-represented among these pairs, albeit only on the side of the larger binding partner. This bias was unique to the native

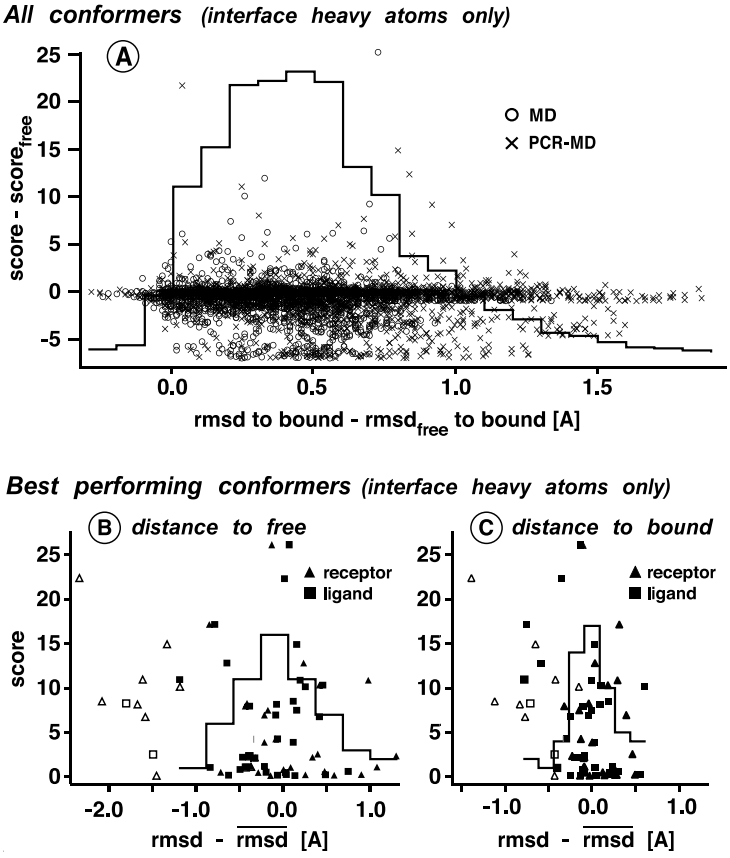

Figure 3.14: Docking performance and structural changes in the interface. In panel (A) the combined distance of the receptor and ligand interface regions from their respective bound conformation is expressed as ($rmsd_{rec}$ + $rmsd_{lig}$)/2 and is plotted against the pair's docking performance. Both values are given relative to the docking of the free conformation pair. Data is shown for each combination of receptor and ligand conformers (11 x 11 x 34). A solid line describes the distribution of rmsd values (distances to the bound structure). Panels (B) and (C) show only the best performing pairs of each ensemble docking. The rmsd of the receptor (triangle) and ligand (square) interface to the free (B) and to the bound structure (C) is given relative to the respective average value of the 10 simulation-derived conformers. High performing conformations seem to be shifted both toward the bound and the free structure. This trend is largely caused by free (experimental) structures (open symbols) that are over-represented on the receptor side of high-performing conformer pairs. Free structures are excluded from the distribution of rmsd shifts (solid lines).

orientation and absent from the conformer pairs with the highest similarity to a random reference (data not shown). Compared to the average ensemble member, experimental conformations (open symbols in figure 3.14B and C) are also closer to the bound structure since the ensembles were moving away from the free without systematically moving toward the bound conformation. The short simulation time sometimes aggravates the effect as it may cause uneven sampling of the conformational space around the starting structure. The preference of experimental receptor structures might be an artifact of the docking protocol being optimized for free and bound crystallographic rather than simulation derived structures – not only in general but actually using the very same test complexes. After excluding experimental structures from the conformations of best complementarity no obvious trend remains, neither to the free experimental nor to the bound state (histogram in figure 3.14B and C).

Indeed, the systematic dependency on a single, for example bound, recognition conformation would impede fast binding. Protein structures move on a flat energy landscape that probably requires milliseconds or even seconds for adequate sampling (Brooks III et al. 1988). The time window for recognition is short by comparison (Northrup and Erickson 1992; Janin 1997; Camacho et al. 2000). Nevertheless, we often observe deviations between the experimental free and bound structures that can only be bridged by large scale correlated motions, which, in turn, are unlikely to occur spontaneously within this short recognition time.

Our extensive data show that short range forces can drive recognition even where this is not evident from the free structures. Due to the simplistic energy function used we can only speculate that the conformations of highest complementarity are related to actual recognition conformers. Our results nevertheless suggest that different such conformers coexist and can be sampled within the short window of opportunity. The cross-docking of simulation-derived structure ensembles indicates that shape-driven recognition does not, or at least not generally, depend on systematic transitions from free to bound structures. This allows us to refine and combine the current models of the protein-protein binding process.

3.4.7 An ensemble model of flexible recognition

Gabdoulline and Wade (2002) recently criticized the mutual inconsistency of current models for protein-protein association. Disputed are the nature of the rate-limiting step (diffusion or induced fit), the shape of the association energy landscape (broad funnel or tight channel), and the mechanism of conformational changes (preexisting equilibrium or induced fit). Most of these inconsistencies can be resolved if we describe binding as a 3-step process of diffusion, free conformer selection, and refolding or "induced fit", as shown in figure 3.15.

Association starts with the diffusional encounter of the two free structure ensembles (R_f and L_f) which, at rate k_1, leads to a micro-collision with approximately correct orientation of receptor and ligand ($R_f \cdots L_f$). The lifetime of this aligned encounter complex allows for gradual desolvation and it could, potentially, be prolonged by random complementarities between sub-populations of the two structure ensembles. Apart from such an unspecific "pre-selection", the structure of the two proteins is still characterized by their free conformation ensembles. This is the point where short range forces and internal dynamics become important for recognition. Specifically matching conformations will select each other from the free conformation ensembles of the two proteins and form a recognition complex ($R_f^* L_f^*$). The recognition complex will quickly be stabilized by progressive desolvation as well as short range electrostatic and van der Waals interactions. At this stage the receptor and ligand structure cannot any longer be considered independent. They are now moving in concert through a potential that has changed from the free to the bound energy landscape. The equilibration into this new landscape requires the transition from the (free) recognition conformations to the more dominant states of the bound structure ensemble ($R_b L_b$). This is potentially a time consuming step, depending on the distance between free and bound structure (or the probability of the recognition conformations in the context of the bound energy landscape) and may be considered a folding process.

In figure 3.15 we attempt to give a schematic view on the free energy profile and the forces that are involved, and compensate each other, at the proposed stages of protein-protein association. This reaction scheme extends earlier 3- and

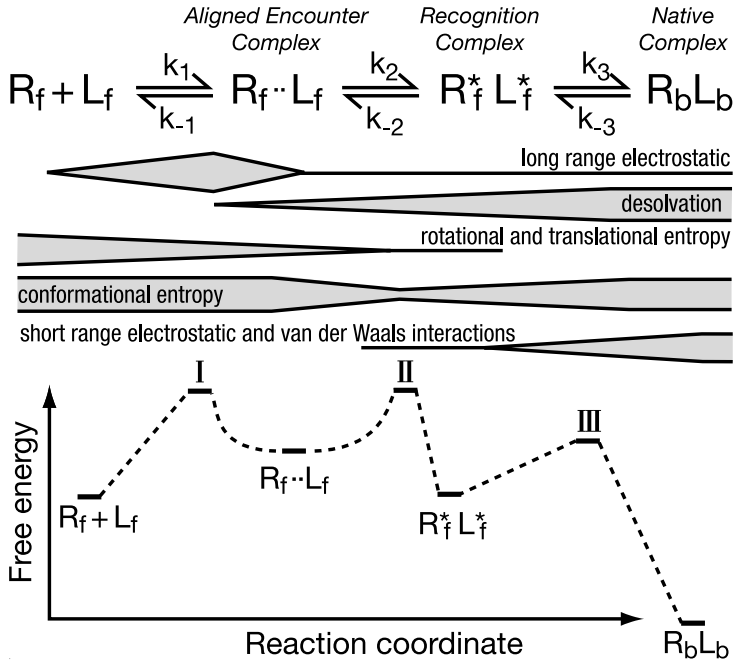

Figure 3.15: A working model for flexible protein recognition. Protein-protein association may be governed by diffusion, selection of matching conformers, and refolding. R_f and L_f are the free structure ensembles of receptor and ligand, respectively. R_f^* and L_f^* are sub-sets of the free receptor and ligand ensembles (recognition conformers). The middle and lower section of the figure suggest, schematically, the forces involved at the different stages and the resulting free energy profile. The widths and barrier heights are not meant to reflect real proportions.

4-state models (Camacho et al. 2000; Frisch et al. 2001; Schreiber 2002) and combines them with the idea of conformer selection (Monod et al. 1965; Kumar et al. 2000; Gabdoulline and Wade 2001). Existing 4-state models (Camacho et al. 2000; Schreiber 2002) distinguish between the formation of an unspecific (randomly aligned) encounter complex on one side and its correct orientation on the other. For the sake of clarity, we combine these two steps into one. The search for this correctly aligned encounter complex ($R_f \cdots L_f$) was considered the rate limiting barrier in the previous models. We introduce an additional step of free conformer selection that separates the diffusive search for a correct orientation from the conformational search for the bound state. Both diffusive and conformational search are well studied in isolation - the former by simulations and experiments on diffusion-controlled associations (Gabdoulline and Wade 2002) and the latter by decades of research on protein folding (Dill and Chan 1997). Conformer selection has been observed in experiments (e.g. (Lancet and Pecht 1976; Foote and Milstein 1994; Leder et al. 1995; Berger et al. 1999)) and our results suggest the specific recognition via a subset of free conformations. Moreover, the mechanism does not rely on the ad-hoc assumption of preexisting bound conformations and is compatible with the time scale and typical rates of protein-protein association.

The scheme contains the previous models as border cases among several possible kinetic regimes: If the free energy cost of selecting matching conformers is much lower than the cost of finding the correct orientation ($k_1 \ll k_2$), the model reverts to the previous 3- or 4-state descriptions (with- or without induced fit, respectively) of a diffusion-controlled reaction. If, on the other hand, we assume that recognition requires bound conformers, the refolding barrier (III in figure 3.15) would be absent ($k_2 \ll k_3$) and we would revert to the preexisting equilibrium model. The proposed 3-step model is the general description of an interaction that can be diffusion controlled, recognition controlled, refolding controlled, or be influenced by a mixture of the three rates.

3.4.8 Implications of the model

Diffusion-controlled associations have been studied experimentally and relative rates for a given system under different conditions can in many cases be reproduced by Brownian Dynamics simulations (Gabdoulline and Wade 2002). An issue with simulations is that association rates are usually overestimated, even if binding is assumed only for orientations very close to the native. Gabdoulline and Wade (2001) showed that this overestimation was different for 5 different protein complexes and concluded that association can be influenced by non-diffusive effects. For the binding of fasciculin-II to acetylcholinesterase in particular, they suggested a mechanism of "conformal gating" by two distinct conformations of a loop.

Our working model of diffusion, selection and refolding offers a similar, more general explanation. The recognition barrier (barrier II in figure 3.15) differs from the free energy of the encounter complex ensemble $R_f \cdots L_f$ by a loss of conformational entropy because it can only be crossed by a sub-set of free conformations. A mixed control by diffusion and recognition should lower observed association rates by a systematic factor (related to the frequency of recognition conformers), such as described by Gabdoullin and Wade. Predominant control by recognition and/or refolding, on the other hand, would uncouple the observed rate from conditions like ionic strength, charge, and viscosity – which they demonstrated for another of the tested complexes.

The 3-step model also helps to refine our description of the transition state ensemble(s) in protein-protein association. Both theoretical (Janin 1997; Camacho et al. 2000) and experimental studies (Frisch et al. 2001) conclude that the transition state closely resembles the structure of the final complex. Less clear is whether or not desolvation is necessary for recognition. According to Camacho et al. (1999, Camacho et al. (2000) partial desolvation is important for the correct positioning and initial stabilization of the encounter complex. However, Frisch et al. (2001) measured activation entropies close to zero for the association of barnase and barstar. They hence assumed that the activated complex remains mostly solvated. This discrepancy may testify to a "special" nature of the barnase

– barstar interface (featuring many charged residues and structural waters). It may on the other hand also result from underlying conformer recognition. Following our 3-step model, recognition occurs at the cost of conformational entropy. A low activation entropy does not rule out desolvation effects but could rather reflect a balance between conformational entropy loss and solvent entropy gain. Our comparison of free and bound MD simulations shows that bound structure ensembles are not generally less diverse than free ones (see 3.3.2 and 3.3.6). One can hence speculate that the refolding phase of binding is accompanied by the re-gain of conformational entropy. A mixed control by diffusion and recognition implies a structurally constrained transition state ensemble that is close to the bound orientation on the one hand, but resembles the two free conformations on the other hand.

3.5 Conclusion

Conformal motions seem to make a considerable impact on the process of protein-protein binding. Already before binding, free interaction interfaces are more flexible than most of the remaining protein surface. Depending on its time scale, the increased mobility may facilitate but also hinder the search for matching conformations and could thus influence the kinetics of protein-protein recognition. Free structure ensembles apparently contain several of these recognition conformations. Recognition, it seems, does not depend on structure transitions from the free to the bound state but occurs between typical members of the free ensemble. I suggest to describe the whole process by a 3-step scheme of diffusion, free conformer selection, and refolding.

The high mobility of binding interfaces could have a negative impact on the thermodynamics of binding. Indeed, the systematic comparison of free and bound structure ensembles shows that binding interfaces loose conformational freedom upon formation of the complex. Nevertheless, in the majority of cases, binding does not restrict a protein's overall motion. Often, the complex gains flexibility in other places. The thermodynamic effect of this loss, gain or redistribution of motion remains difficult to estimate. Entropy calculations are hindered by insufficient

sampling of conformational space and turn out to be distorted by spurious correlations between deterministic simulations. The latter artifact can be eliminated and I calculated the entropy differences for a subset of protein complexes. During binding, conformational entropy can be both lost or gained. In most cases, this entropic contribution should have an important effect on the overall free energy of binding.

Structure fluctuations should thus exert influence both on the speed of recognition and the thermodynamic stability of protein-protein complexes. This interplay of protein flexibility and recognition remains far from understood. Most theoretical but also experimental studies have so far focused on diffusion-controlled interactions without large changes in the binding interface. It is now time to move on to systems where association is either depending on or followed by large-scale structural rearrangements.

3.6 Methods

3.6.1 Short conformational sampling

Simulations were performed with a modified version of X-PLOR (Brünger 1992; Abseher and Nilges 2000) using the CHARMM19 force field (Brooks et al. 1983) and an electrostatic cutoff of 12 Å with force shifting (Steinbach et al. 1991).

The coordinates of the 51 molecules (table 3.1) were retrieved from the Protein data bank (Berman et al. 2002). An automated procedure removed duplicate peptide chains and all hetero atoms (but not waters), converted non-standard amino acids to their closest standard residue and identified disulfide bonds. Missing atoms, including polar hydrogens were added and briefly minimized. The protein was surrounded by a 9 Å layer of TIP3 water molecules and the solvent briefly equilibrated. 10 copies were starting point for parallel simulations of 50 ps length summing up to 500 ps total simulation time per system. SHAKE constraints (van Gunsteren and Berendsen 1977) were put on all bonds to hydrogens and on all TIP3 waters. Each copy was heated from 100 K to 300 K in 50 K steps of 1 ps

each, followed by additional 5 ps of equilibration with continued re-assignment of velocities every 1 ps. The temperature was kept constant by explicit coupling to a heat bath via Langevin dynamics and a friction coefficient of 20 ps^{-1} for water oxygens and between 0.5 and 5.5 ps^{-1} for protein atoms dependent on their solvent accessible area. A time step of 2 fs was used. The simulation scripts are available upon request.

A second set of simulations was performed with identical setup but adding an additional force onto the potential acting along the principle components of motion, basically as described by Abseher and Nilges (2000). In difference to the published method we re-defined the principal components iteratively during the calculation.

For the docking ensembles (section 3.4), 100 snapshots spaced 5 ps apart were taken from each 10 trajectories. The snapshots were fitted to their average structure and divided into 10 groups by c-means fuzzy clustering (Gordon and Somorjai 1992) over the coordinates of backbone carbonyl carbon and every second side chain carbon. The clustering method is similar to the simple k-means but gives each item a continuous membership to each cluster instead of a binary membership to one. From each cluster the structure nearest to the center was selected for docking.

3.6.2 Extended conformational sampling

Extended simulations were performed with the Amber 7.0 program package using the modified all-atom force field parm98 (Cornell et al. 1995; Kollman et al. 1997). They were based on the automatically processed structures described above. The following protocol was consistently applied to all 21 protein structures. Hydrogens and waters were stripped off, and the proteins subjected to a WhatIf hydrogen bond network optimization (Vriend 1990) that adjusted the protonation state of histidines and flipped carboxylamide moieties of certain Glutamine and Asparagine residues. Chain breaks and premature ends of peptide chains (due to unresolved residues in the original structure) were capped with a N-methylamine or Acetyl group but no attempts were made to model missing

residues. Hydrogens were added with the tleap program (part of the Amber suite). The protein was surrounded by a rectangular box of TIP3P water models keeping a minimal distance of 10 Å between protein and border of the solvent box. Any net charge of the protein was neutralized with Sodium or Chlorine ions. The solvent was minimized while keeping all protein coordinates restrained.

The following simulation protocol was applied to 10 independent copies of the system. Bond lengths involving hydrogen atoms were fixed with the SHAKE algorithm (van Gunsteren and Berendsen 1977). Periodic boundary conditions were applied with a direct-space non-bonded cutoff of 9 Å and particle mesh Ewald (PME) treatment of long-range electrostatic forces (Essmann et al. 1995). The solvent was heated to 300 K over a 10 ps constant volume MD and equilibrated with a 10 ps constant pressure MD at 300 K keeping the protein coordinates harmonically restrained (K = 50 kcal $mol^{-1}Å^{-1}$) and applying an integration time step of 1 fs together with temperature coupling to a heat bath (time constant 0.5 ps) (Berendsen et al. 1984). Restraints on the solute were then stepwise released during a final 20 ps constant pressure MD. Over the whole 40 ps equilibration phase, temperatures were reassigned every 1.01 ps from a Maxwell distribution. A production MD of 1 ns was then performed using an integration time step of 2 fs under NVT conditions at 300 K controlled by the Berendsen coupling algorithm with the default time constant of 1 ps. Structure snapshots were saved every 100 fs. The complete protocol of minimization, equilibration and simulation was automated, parallized and applied in identical fashion to all 21 proteins.

This resulted in 10 single 1 ns trajectories for each free receptor, free ligand and complex, respectively. I calculated the trace of mean C_α distance to the last structure for each single 1 ns simulation and determined the gradient of this distance over the last 500 ps by a linear least-squares fit (excluding the last 50 ps). I classified single trajectories as (non-equilibrated) ”outliers” if their gradient fell 1.5 standard deviations below the average of all 10 simulations.

3.6.3 Surface patches

All atoms not present in both free and bound receptor or ligand structure were removed before performing the analysis. WhatIf was used to calculate the accessible surface area of each atom relative to the total exposure possible (Vriend 1990). The protein surface was defined as any atom with more than 5% relative accessible surface area (in the free structure). The binding patch was defined as any atom within 6 Å of the other molecule (in the structure of the complex). Random surface patches comprising the same number of atoms as the binding patch (n) were generated by picking a random surface atom and assigning the n nearest surface atoms to the new patch. The patch was discarded if more than 25% of its atoms overlapped with another random or the binding patch. The procedure was repeated 50 times and yielded different numbers of random patches, depending on the size of the binding patch and the remaining surface.

3.6.4 Flexibility

Flexibility was defined as the average pairwise distance between simulation snapshots. Snapshots were extracted from the last 30 ps of the 10 x 50 ps simulations in an interval of 2 ps. I then calculated the root mean square (rms) distance (after individual least-squares fitting) between every pair of structures that did *not* stem from the same 50 ps trajectory. The flexibility of a certain protein or part of a protein is thus the average of 20250 rms values between 150 simulation snapshots. The same procedure was applied to the 10 x 1 ns simulations. Here, snapshots were taken in an interval of 5 ps from the last 500 ps of each 1 ns trajectory. Water molecules, hydrogens and any atom not present in both free and bound structure were removed prior to the analysis. The calculation of pairwise distances was parallized and distributed to between 30 and 90 processors of a Linux cluster.

3.6.5 Entropy calculations

Entropy differences were determined from a combination of several quasiharmonic calculations applied to both free and bound trajectories using different pro-

tocols for the superposition and re-arrangement of coordinate frames. Each single protocol consisted of the following steps: (1) Adjusting the number of trajectories: Certain single trajectories were excluded from the calculation either because they were classified as outliers (see section 3.6.2) or in order to estimate errors. Further trajectories were removed arbitrarily so as to adjust receptor, ligand, and complex ensembles to the same number of independent trajectories. In most cases, 9 trajectories were used. (2) Adjusting the atom content: I removed all hydrogens and non-protein atoms, as well as any atom that was not present in both free and bound structures. (3) Adjusting the frame order: For the analysis of real and spurious correlations across the binding interface, receptor coordinates were in some protocols paired up with ligand coordinates from an independent trajectory. The relative time order of ligand frames was retained or shuffled. (4) Iterative rms-fitting: The single trajectories were iteratively fitted to their respective average until the rms distance between the last and the previous average structure fell below 10^{-6} Å and then transformed "en-block" onto the bound state. To shorten the number of iterations, coordinate frames were initially superimposed onto the experimental starting structure. (5) The modified set of coordinate frames was exported into Amber file format and passed to the ptraj program for the calculation of the mass-weighted covariance matrix and determination of vibrational, rotational and translational entropies. I automated steps (1) through (5) and parallelized the calculation of the different protocols.

A set of complete and disassembled entropy changes such as given in table 3.3 required the analysis of free and bound trajectories with 10 protocols. The protocols differed in (1) the source of the coordinate data i.e. simulation of free receptor, free ligand, or complex; (2) Which part of the structure was analyzed i.e. receptor, ligand, or both; (3) The superpositioning of receptor and ligand frames i.e. separately or as single molecule; (4) Whether receptor and ligand data were from the same (complex) trajectory or from independent simulations; (5) The order of ligand frames i.e. left intact or shuffled with respect to the receptor data. The 10 protocols are summarized in table 3.2. The change of complete vibrational entropy was determined by comparing the entropies calculated with protocols 'com' and 'fcom' (see table 3.2).

For each complex, I evaluated the convergence of free vs. bound entropy differences by repeating the 10 calculations using different starting frames and different frame offsets. Errors were estimated by repeating all calculations 6 times while excluding 3 different single trajectories from the ensembles of receptor, ligand, and complex simulations. For example, an absolute entropy calculated from the coordinates of 9 independent 1 ns trajectories was recalculated after removing trajectories (1,2,3), (4,5,6), (7,8,9), (1,4,7), (2,5,8) or (3,6,9). The standard deviation σ_Δ for the difference of two absolute entropies with standard deviations σ_1 and σ_2 was then calculated as $\sigma_\Delta = (\sigma_1^2 + \sigma_2^2)^{1/2}$. Variances stem thus from calculations on less data than used for the reference value (e.g. 6x1 ns instead of 9x1 ns) and constitute a conservative (high) estimate.

3.6.6 Docking

All protein-protein docking calculations were performed with HEX version 4.2 (Ritchie and Kemp 2000). Orientations were discriminated by shape complementarity only. For all protein independent parameters the default values provided by HEX were used with the exception of the distance range step, which was set to 0.5 Å and the receptor and ligand samples which were set to 720 (Ritchie and Kemp 2000). The initial molecular separation and the distance range to be sampled were calculated from the maximal and minimal distance from the center of mass to any surface atom (any atom with an exposure >95% as determined by WhatIf (Vriend 1990)). In the 7 cases where the receptor had a radius larger than 35 Å HEX "macro docking" was performed with default parameters, i.e. the program docked the ligand to several overlapping fragments of the receptor (Ritchie 2003). The 512 highest scoring solutions were retained from each docking, thus the combination of 11 x 11 conformations always produced 61952 orientations. The docking of a single conformer pair took in the order of 15 min on a dual 2.4 GHz Xeon computer but lasted about 8 h for the "macro docking" cases.

3.6.7 Randomized reference complexes

For each ligand we generated 100 transformation matrices with randomized Euler angles and a random translation onto a sphere around the receptor's center of mass. These randomized orientations were each subjected to 100 steps of rigid body minimization using a soft van der Waals potential and a NOE restraint pulling the two centers of mass together (X-PLOR script available upon request). We removed all orientations having any atom contact in common with the native complex (fnac > 0) and performed a hierarchical clustering by the pairwise overlap of atom contacts. The clustering will be described in detail elsewhere. For the present purpose, we applied a clustering threshold of 0.0001 and obtained a set of cluster centers without mutual contact overlap. We selected 10 at random and re-calculated the "fnac" of all HEX solutions with respect to each of the 10 random complexes. From these values we estimated the probability of the score (for reproducing the native complex) being a random observation. The necessary random distribution cannot be deduced from 10 values. However, score values were by definition positive and usually small. A lognormal distribution was hence the least biased assumption.

3.6.8 Specificity estimate for docking scores

Based on 10 random scores and the assumption that docking scores follow a lognormal distribution, the "specificity" of a docking result can be estimated. The following solution to this problem was worked out by Michael Habeck and Wolfgang Rieping. Given is the score s for the success of a docking experiment to reproduce the native complex and the scores r_1 .. r_{10} to reproduce 10 non-native random complexes. We assume random scores to follow a lognormal distribution. The lognormal density function $f(x)$ can be estimated from the mean α and the standard deviation β of the 10 log-transformed random scores.

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \ln r_i \tag{3.3}$$

$$\beta = \sqrt{1/(n-1)} \sum_{i=1}^{n} (\ln r_i - \alpha)^2 \tag{3.4}$$

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp(-\frac{1}{2}[\frac{\ln x - \alpha}{\beta}]^2) \tag{3.5}$$

Given $f(x|\alpha,\beta)$, we determined the confidence level κ of the smallest interval still containing s.

$$\kappa = \int_{r_{max}^2/s}^{s} f(x|\alpha,\beta)dx \tag{3.6}$$

$$r_{max} = \exp(\alpha - \beta^2) \tag{3.7}$$

$$\kappa = \frac{1}{2}\mathrm{erf}(\frac{\ln s - \alpha}{\sqrt{2\beta^2}}) - \frac{1}{2}\mathrm{erf}(\frac{\ln(r_{max}^2/s) - \alpha}{\sqrt{2\beta^2}}) \tag{3.8}$$

κ is the probability that s is not a random observation such as r_1 .. r_{10}.

3.6.9 Miscellaneous analysis

All atoms not present in both free and bound receptor or ligand structure were removed before performing any comparative analysis. The binding interface of a protein was defined as any atom within 6 Å of the other binding partner. This gives patches of similar size as the more traditional residue based definition that considers any residue having, at least, one atom within 4.5 Å from the other molecule. For consistency with docking studies the latter definition was used in section 3.4. Many calculations (including docking and analysis) were automated and distributed to between 30 and 90 processors of a Linux cluster. The computation time spent for this chapter would translate, very roughly, to more than 25 years on a single 2.4 GHz CPU.

3.6.10 Figures

Figure 3.1 was prepared with VMD (Humphrey et al. 1996). Figure 3.10 was prepared with MOLMOL (Koradi et al. 1996). Figures 3.2 through 3.9, as well as 3.11 and 3.12 were created using Biggles (biggles.sourceforge.net). Figures 3.13 and 3.14 were created with IgorPro (www.wavemetrics.com). For figure 3.7, I made use of an extension to Biggles, programmed and kindly provided by Wolfgang Rieping. All figures of section 3.4 were devised together with – and ultimately prepared by Johan Leckner.

Chapter 4

Conclusion

4.1 Proteins on the edge

The unfolding of a single molecule and the recognition between two proteins are two rather different events. However, both processes push a complex system of interacting atoms from one conformation ensemble/state to another. This reminds to some extend of certain ”phase transitions” described by statistical physics although the concept does not apply in a strict sense (Ball 1999). Second order phase transitions, for example the change from a liquid / gas two-phase system to a single fluid phase, proceed through a singularity, the ”critical point”, where the many-body system of interacting particles complies with neither of the two states. In the vicinity of this point, microscopic fluctuations have increasing influence on macroscopic properties until the system is on all scales abandoned to this random variation. Proteins between free and bound or folded and unfolded state are likewise ”on the edge”. Benign thermal fluctuations may thus exert major effects on the whole process (and the parallel with phase transitions may or may not end there).

However, for a protein, ”being on the edge” does not constitute an anomaly. Life itself operates off equilibrium and a cell or organism at thermodynamic equilibrium is dead. On the one hand, most proteins may move most of the time through some equilibrated structure ensemble – bound or free, active or inactive,

waiting for their signal, target, substrate or destruction. On the other hand, this state may be scientifically rather irrelevant; Function is defined in action, when a protein is recognizing, catalyzing, switching, folding or unfolding. The previous chapters have examined two different such processes and structure flexibility turns out to play a decisive role in both.

Chains of spectrin repeats apparently rely on structure flexibility to achieve a smooth response to external force. Single molecule experiments on this domain, in accord with simulations, show clear traces of structure fluctuation and this picture is supported by experiments on mutated repeats. On the verge of disruption, thermal fluctuations decide how much extension a spectrin repeat tolerates and whether or not unfolding is blocked by intermediate non-native structures.

Structure fluctuations may both complicate and facilitate the interaction between two proteins. In many of the cases that I studied, only a subset of the two free structure ensembles was mutually compatible. A conformer selection step may thus impede the rate of recognition. Interaction sites turned out to be more flexible than normal protein surfaces, which might hinder this selection further. However, recognition most likely requires only an approximate fit. Increased flexibility may as well facilitate the selection of recognition conformers within the short time of a protein collision. Somewhat surprisingly, the formation of a complex did not necessarily restrict the diversity of structure ensembles. Overall conformational entropy could rise as well as fall. Even the entropic cost of fixing one protein to another was in part recovered by new motions between the binding partners. Thus, structure fluctuations exert a considerable influence on both the formation and the stability of protein-protein complexes.

4.2 Next?

The results and ideas presented throughout my thesis are incomplete in several respects and leave many questions unanswered.

The atomic force microscopy experiments of my collaborators provide only indirect support to the detailed pathways that I proposed for the unfolding of spec-

trin repeats. Convincing evidence would require further experiments. Recently, Sarkar et al. (2004) succeeded to simultaneously measure fluorescence and force response of a single molecule. This could put a new generation of experiments into the realm of possibility, that combine atomic force microscopy with the measurement of (distance dependent) fluorescence resonance energy transfer. Such a simultaneous observation of intra-domain distances and unfolding lengths and forces would be a very elegant way to prove or disprove the suggested pathways and intermediates of spectrin unfolding. The role of pre-stretching and structure diversity for the overall elasticity of spectrin networks could perhaps be examined with similar experiments that monitor the distance between neighboring repeats in the unfolding spectrin chain. Moreover, it would be interesting to study the function of spectrin repeats in proteins other than spectrin itself. Finally, with some imagination, spectrin repeats might even find applications as molecular shock absorber in future nanotechnological devices.

Also our understanding of protein-protein interaction would benefit from further experiments and simulations. Most of the complexes, for which we have detailed structural data (of free and bound state), lack any information about the kinetics and thermodynamics of binding. Theoreticians and experimentalists often work on different proteins and rarely reconcile their results. Joint efforts should therefore concentrate on a common set of interactions. This set should be as diverse as possible and should also include complexes that exhibit large structural rearrangements upon binding. Another priority could be the development of single molecule experiments that aim at the complexities of structure fluctuations during protein-protein recognition. It remains to be seen to which extent structure fluctuations indeed influence the kinetics and thermodynamics of binding and whether the three-step model proposed in section 3.4.7 constitutes a suitable framework for these studies. A related question concerns the level of detail that is necessary for recognition – how close has to be the match between recognition conformers and to which extent is conformer selection also a cooperative process, more in line with the classic idea of induced fit?

Better representations of protein dynamics are obviously necessary for the prediction of protein complexes from separate structures. Johan Leckner and I are

currently experimenting with protein flexibility at several stages of a protein docking pipeline. Section 3.4 already described our strategy to extend the sampling of orientations by a combinatorial cross-docking of simulation snapshots. We are also testing the effect of short molecular dynamics simulations on the refinement and evaluation (scoring) of docking solutions.

Protein interaction networks receive a large share of the renewed interest in "systems biology" (Kitano 2002) and it's not all buzz – we have seen remarkable advances in the large scale detection of such interactions (Gavin et al. 2002). At the same time, structural genomics efforts turn out ever increasing numbers of independent protein structures (Zhang and Kim 2003). Efficient methods and sound theories for the structural study of protein interactions are, evidently, in high demand.

4.3 Complexes of complex molecules in complex cells of complex organisms in a complex environment

Mache die Dinge so einfach wie möglich – aber nicht einfacher.[1]

Albert Einstein

Life seems like a rather complicated matter. Every living cell is built from many thousands of different molecules that interact and create and modify each other. Every human being comprises billions of cells that somehow self-organize into tissues and repair, destroy and nurture one another. Human beings, as any other animal, are themselves embedded in a web of dependencies, competition, and cooperation with their living environment. On each of these layers we see complex traits and sophisticated behavior emerging from interactions between simpler entities.

Underlying complexity is often most apparent when the system moves from one state to another. The study of growth and development tells more about cellular interactions and signals than the momentary layout of the adult organism.

[1] Make everything as simple as possible – but not simpler.

Static "wiring diagrams" of biochemical pathways are informative but not sufficient for proper understanding of metabolism. Long standing textbook knowledge about the regulation of glycolysis, for example, is disproved by the analysis of the metabolic flux through this cascade of enzymes (Teusink et al. 1998; Bakker et al. 1999). Gene regulatory networks illustrate the same point. The low concentration of transcription factors renders genetic regulation prone to stochastic fluctuations (McAdams and Arkin 1999). Regulatory circuits have to suppress this noise but can, in some instances, also utilize it to achieve variability in a population of cells. Random concentration fluctuations may thus decide over the fate of a cell on the edge between two genetic programs (McAdams and Arkin 1999).

In this work, I have examined fluctuations at the level of atoms in single molecules. Parallels appear between effects of structure dynamics and the role of concentration fluctuations one detail level further up. Just like cells need to suppress stochastic concentrations, proteins need to recognize each other in spite of atomic fluctuations. And like cell diversity may be based on concentration noise, the elasticity of the membrane skeleton may, to some extent, go back to thermal fluctuations of protein structure.

As Richard Feynman put it, "everything that living things do ... can be described as the wiggling and jiggling of atoms." Will we hence need to know the position and vibration of each atom before we can understand and predict the behavior of a cell? Would this knowledge actually make us understand? The answer to both question is probably No. Life has evolved to be robust to noise and change and it has evolved to evolve. This process has shaped functional modules into hierarchical layers of complexity like, for example, protein domains, gene cassettes, cells, and organisms (Csete and Doyle 2002). Lower level detail will probably not be relevant for many higher level functions. Yet, at the moment of perturbation or transition, the hidden complexity of lower layers may turn into a dominating force. Hence, we need to study and understand complexity at each level of detail. We need to generalize our complicated results and translate them into trends. The challenge will be to integrate this knowledge and to find the level of detail that is relevant for a given question. The jiggling and wiggling of atoms in proteins has to be part of this global picture and certainly holds some surprises for future research.

Acknowledgments

Many people have contributed to this work in many different ways. Wherever possible, I tried to pay due credit directly in the text. However, often aid cannot be quantified in figures, plots or algorithms.

First of all, I would like to thank Michael Nilges for taking me under his wings and for giving me the freedom and means to act on my curiosity but also for offering (not only scientific) advise and support whenever needed.

I am indebted to Jeremy Smith, who readily accepted to supervise me as external PhD student and who provided valuable advise and helped in many ways in a straightforward manner.

I am grateful to the Boehringer Ingelheim Fonds for the generous financial support of my work. The friendly B.I.F. team was caring and always very helpful.

My very special thanks go to Johan Leckner. Together we worked through ten thousands of lines of python code, swore at as many bugs, got lost in terabytes of data, and pursued untold trends with countless plots. Without him this whole enterprise wouldn't have come half that far and wouldn't have been a tenth the fun. Also the two other members of our little expatriate cell cannot go unmentioned: Michael Habeck and Wolfgang Rieping supplied crucial bits of code, cleared up complicated problems, and, over one or two (or three) cups of coffee, were always willing to enlighten Johan and me on tricky issues of physics and statistics.

However, any enlightenment wouldn't have carried far, if not for Tru Huynh, who ceaselessly replaced burnt processors, failed memory chips or overworked motherboards and patiently satisfied even the most exotic software wishes. I also want to thank Renée Communal who gave (badly needed) rear cover in the everyday skirmish with bureaucracies and travel agents.

Many other past and present members of the Nilges group made this work a pleasant experience and new friends in Heidelberg as well as Paris took care that there was a life beyond work.

Last, and anything but least, I thank my parents for their loving encouragement and unconditional support over all these years.

Publications

The work presented in this thesis was (or will be) described in the following articles:

- Altmann S.M.*, R. Grünberg*, P.F. Lenne, J. Ylänne, A. Raae, K. Herbert, M. Saraste, M. Nilges, J.K. Hörber (2002): Pathways and intermediates in forced unfolding of spectrin repeats. *Structure* 10, 1085-96.
- Grünberg, R.*, J. Leckner*, M. Nilges (2004): Complementarity of structure ensembles in protein-protein binding. *Structure* 12, 2125-36.
- Grünberg, R., M. Nilges, J. Leckner (2005): The dynamics of protein-protein binding. *in preparation.*

(* shared first co-authors)

References

Abagyan, R. and M. Totrov (1997). Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J Mol Biol 268*(3), 678–85.

Abseher, R. and M. Nilges (2000). Efficient sampling in collective coordinate space. *Proteins 39*(1), 82–8.

Altmann, S., R. Grünberg, P. Lenne, J. Ylänne, A. Raae, K. Herbert, M. Saraste, M. Nilges, and J. Hörber (2002). Pathways and intermediates in forced unfolding of spectrin repeats. *Structure (Camb) 10*(8), 1085–96.

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science 181*(96), 223–30.

Ansari, A., J. Berendzen, S. Bowne, H. Frauenfelder, I. Iben, T. Sauke, E. Shyamsunder, and R. Young (1985). Protein states and proteinquakes. *Proc Natl Acad Sci U S A 82*(15), 5000–4.

Arold, S., R. O'Brien, P. Franken, M. Strub, F. Hoh, C. Dumas, and J. Ladbury (1998). RT loop flexibility enhances the specificity of Src family SH3 domains for HIV-1 Nef. *Biochemistry 37*(42), 14683–91.

Arumugam, S., G. Gao, B. L. Patton, V. Semenchenko, K. Brew, and S. R. Van Doren (2003). Increased backbone mobility in beta-barrel enhances entropy gain driving binding of N-TIMP-1 to MMP-3. *J Mol Biol 327*, 719–734.

Ashkin, A. (1987). Optical trapping and manipulation of viruses and bacteria. *Science 235*(4795), 1517–1520.

Astumian, R. and M. Bier (1994). Fluctuation driven ratchets: Molecular motors. *Phys Rev Lett 72*(11), 1766–1769.

Austin, R., K. Beeson, L. Eisenstein, H. Frauenfelder, and I. Gunsalus (1975). Dynamics of ligand binding to myoglobin. *Biochemistry 14*(24), 5355–73.

Bader, G. and C. Hogue (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol 20*(10), 991–7.

Bakker, B., P. Michels, F. Opperdoes, and H. Westerhoff (1999). What controls glycolysis in bloodstream form Trypanosoma brucei? *J Biol Chem 274*(21), 14551–9.

Ball, P. (1999). Transitions still to be made. *Nature 402*, C73–C76.

Balsera, M., S. Stepaniants, S. Izrailev, Y. Oono, and K. Schulten (1997). Reconstructing potential energy functions from simulated force-induced unbinding processes. *Biophys J 73*, 1281–1287.

Balsera, M., W. Wriggers, Y. Oono, and K. Schulten (1996). Principal component analysis and long time protein dynamics. *J Phys Chem 100*, 2567–72.

Bashford, D. and D. Case (2000). Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem 51*, 129–52.

Berendsen, H., J. Postma, W. Van Gunsteren, A. DiNola, and J. Haak (1984). Molecular dynamics with coupling to an external heat bath. *J Chem Phys 81*, 3684–3690.

Berger, C., S. Weber-Bornhauser, J. Eggenberger, J. Hanes, A. Pluckthun, and H. Bosshard (1999). Antigen recognition by conformational selection. *FEBS Lett 450*(1-2), 149–53.

Berman, H., T. Battistuz, T. Bhat, W. Bluhm, P. Bourne, K. Burkhardt, Z. Feng, G. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. Westbrook, and C. Zardecki (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr 58*(Pt 6 No 1), 899–907.

Best, R., B. Li, Steward, V. Daggett, and J. Clarke (2001). Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophys J 81*, 2344–2356.

Betts, M. and M. Sternberg (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng 12*(4), 271–83.

Bhat, T., G. Bentley, G. Boulot, M. Greene, D. Tello, W. Dall'Acqua, H. Souchon, F. Schwarz, R. Mariuzza, and R. Poljak (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc Natl Acad Sci U S A 91*(3), 1089–93.

Binnig, G., C. Quate, and C. Gerber (1986). Atomic force microscope. *Phys Rev Lett 56*(9), 930–933.

Bosshard, H. (2001). Molecular recognition by induced fit: how fit is the concept? *News Physiol Sci 16*, 171–3.

Brady, G. and K. Sharp (1997). Entropy in protein folding and in protein-protein interactions. *Curr Opin Struct Biol 7*(2), 215–21.

Brooks, B., R. Bruccoleri, Olafson B.D., D. States, S. Swaminathan, and M. Karplus (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J Comp Chem 4*, 187–217.

Brooks III, C., M. Karplus, and B. Pettitt (1988). *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*. New York: Wiley.

Brünger, A. (1992). *X-PLOR. A System for X-ray crystallography and NMR*. New Haven: Yale University Press.

Bruschweiler, R. (2003). New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Curr Opin Struct Biol 13*(2), 175–83.

Byers, T. and D. Branton (1985). Visualization of the protein associations in the erythrocyte membrane skeleton. *Proc Natl Acad Sci U S A 82*, 6153–6157.

Camacho, C., S. Kimura, C. DeLisi, and S. Vajda (2000). Kinetics of desolvation-mediated protein-protein binding. *Biophys J 78*(3), 1094–105.

Camacho, C., Z. Weng, S. Vajda, and C. DeLisi (1999). Free energy landscapes of encounter complexes in protein-protein association. *Biophys J 76*(3), 1166–78.

Carrion-Vazquez, M., P. Marszalek, A. Oberhauser, and J. Fernandez (1999). Atomic force microscopy captures length phenotypes in single proteins. *Proc Natl Acad Sci U S A 96*, 11288–11292.

Carrion-Vazquez, M., A. Oberhauser, T. Fisher, P. Marszalek, H. Li, and J. Fernandez (2000). Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog Biophys Mol Biol 74*(1-2), 63–91.

Carrion-Vazquez, M., A. Oberhauser, S. Fowler, P. Marszalek, S. Broedel, J. Clarke, and J. Fernandez (1999). Mechanical and chemical unfolding of a single protein: a comparison. *Proc Natl Acad Sci U S A 96*, 3694–3699.

Case, D. (1994). Normal mode analysis of protein dynamics. *Curr Opin Struct Biol 4*, 285–290.

Caves, L., J. Evanseck, and M. Karplus (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci 7*(3), 649–66.

Chen, R., J. Mintseris, J. Janin, and Z. Weng (2003). A protein-protein docking benchmark. *Proteins 52*(1), 88–91.

Chothia, C. and A. Lesk (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J 5*(4), 823–6.

Cieplak, P., J. Caldwell, and P. Kollman (2001). Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: Aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of nucleic acid bases. *J Comput Chem 22*, 1048–1057.

Cole, C. and J. Warwicker (2002). Side-chain conformational entropy at protein-protein interfaces. *Protein Sci 11*, 2860–2870.

Cornell, W., R. Abseher, M. Nilges, and D. A. Case (2001). Continuum solvent molecular dynamics study of flexibility in interleukin-8. *J Mol Graph Model 19*, 136–145.

Cornell, W., P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc 117*, 5179–5197.

Craig, D., A. Krammer, K. Schulten, and V. Vogel (2001). Comparison of the early stages of forced unfolding for fibronectin type iii modules. *Proc Natl Acad Sci U S A 98*, 5590–5595.

Csete, M. and J. Doyle (2002). Reverse engineering of biological complexity. *Science 295*(5560), 1664–9.

De Matteis, M. and J. Morrow (1998). The role of ankyrin and spectrin in membrane transport and domain formation. *Curr Opin Cell Biol 10*(4), 542–9.

Dill, K., S. Bromberg, K. Yue, K. Fiebig, D. Yee, P. Thomas, and H. Chan (1995). Principles of protein folding–a perspective from simple exact models. *Protein Sci 4*(4), 561–602.

Dill, K. and H. Chan (1997). From Levinthal to pathways to funnels. *Nat Struct Biol 4*(1), 10–9.

Discher, D. and P. Carl (2001). New insights into red cell network structure and spectrin unfolding - a current review. *Cell Mol Biol Lett 6*, 593–606.

Djinovic-Carugo, K., P. Young, M. Gautel, and M. Saraste (1999). Structure of the alpha-actinin rod: molecular basis for cross-linking of actin-filaments. *Cell 98*, 537–546.

Drewes, G. and T. Bouwmeester (2003). Global approaches to protein-protein interactions. *Curr Opin Cell Biol 15*(2), 199–205.

Ehrlich, L. P., M. Nilges, and R. Wade (2005). The impact of protein flexibility on protein-protein docking. *Proteins 58*(1), 126–133.

Elgsaeter, A., B. Stokke, A. Mikkelsen, and D. Branton (1986). The molecular basis of erythrocyte shape. *Science 234*, 1217–1223.

Engler, N., A. Ostermann, N. Niimura, and F. Parak (2003). Hydrogen atoms in proteins: positions and dynamics. *Proc Natl Acad Sci U S A 100*(18), 10243–8.

Epstein, C., R. Goldberger, and C. Anfinsen (1963). The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harbor Symp Quant Biol 28*, 439–449.

Essmann, U., L. Perera, M. Berkowitz, T. Darden, H. Lee, and L. Pedersen (1995). A smooth particle mesh Ewald method. *J Chem Phys B 103*, 6998–7014.

Fayos, R., G. Melacini, M. Newlon, L. Burns, J. Scott, and P. Jennings (2003). Induction of flexibility through protein-protein interactions. *J Biol Chem 278*(20), 18581–7.

Fenimore, P., H. Frauenfelder, B. McMahon, and F. Parak (2002). Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc Natl Acad Sci U S A 99*(25), 16047–51.

Fenimore, P., H. Frauenfelder, B. McMahon, and R. Young (2004). Bulk-solvent and hydration-shell fluctuations, similar to alpha- and beta-fluctuations in glasses, control protein motions and functions. *Proc Natl Acad Sci U S A 101*(40), 14408–13.

Figeys, D. (2003). Novel approaches to map protein interactions. *Curr Opin Biotechnol 14*(1), 119–25.

Finkelstein, A. and J. Janin (1989). The price of lost freedom: entropy of bi-molecular complex formation. *Protein Eng 3*(1), 1–3.

Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Ber Dt Chem Ges 27*, 2985–93.

Fisher, T., A. Oberhauser, M. Carrion-Vazquez, P. Marszalek, and J. Fernandez (1999). The study of protein mechanics with the atomic force microscope. *Trends Biochem Sci 24*(10), 379–84.

Foote, J. and C. Milstein (1994). Conformational isomerism and the diversity of antibodies. *Proc Natl Acad Sci U S A 91*(22), 10370–4.

Forman-Kay, J. D. (1999). The 'dynamics' in the thermodynamics of binding. *Nat Struct Biol 6*, 1086–1087.

Frauenfelder, H. and D. Leeson (1998). The energy landscape in non-biological and biological molecules. *Nat Struct Biol 5*(9), 757–9.

Frauenfelder, H., S. Sligar, and P. Wolynes (1991). The energy landscapes and motions of proteins. *Science 254*(5038), 1598–603.

Frisch, C., A. Fersht, and G. Schreiber (2001). Experimental assignment of the structure of the transition state for the association of barnase and barstar. *J Mol Biol 308*(1), 69–77.

Frisch, C., G. Schreiber, C. Johnson, and A. Fersht (1997). Thermodynamics of the interaction of barnase and barstar: changes in free energy versus changes in enthalpy on mutation. *J Mol Biol 267*(3), 696–706.

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

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

Gao, J. and D. Truhlar (2002). Quantum mechanical methods for enzyme kinetics. *Annu Rev Phys Chem 53*, 467–505.

Gavin, A., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415*(6868), 141–7.

Gogonea, V., D. Suarez, A. van der Vaart, and K. Merz, Jr (2001). New developments in applying quantum mechanics to proteins. *Curr Opin Struct Biol 11*(2), 217–23.

Goh, C., D. Milburn, and M. Gerstein (2004). Conformational changes associated with protein-protein interactions. *Curr Opin Struct Biol 14*(1), 104–9.

Gohlke, H. and D. Case (2004). Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem 25*(2), 238–50.

Goodford, P. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem 28*(7), 849–57.

Gordon, H. and R. Somorjai (1992). Fuzzy cluster analysis of molecular dynamics trajectories. *Proteins 14*(2), 249–64.

Gray, J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. Rohl, and D. Baker (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol 331*(1), 281–99.

Grum, V. L., D. Li, R. I. MacDonald, and A. Mondragon (1999). Structures of two repeats of spectrin suggest models of flexibility. *Cell 98*, 523–535.

Grünberg, R., J. Leckner, and M. Nilges (2004). Complementarity of structure ensembles in protein-protein binding. *Structure (Camb) 12*(12), 2125–36.

Gsponer, J. and A. Caflisch (2001). Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J Mol Biol 309*, 285–298.

Halperin, I., B. Ma, H. Wolfson, and R. Nussinov (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins 47*(4), 409–43.

Hansen, J., R. Skalak, S. Chien, and A. Hoger (1996). An elastic network model based on the structure of the red blood cell membrane skeleton. *Biophys J 70*, 146 –166.

Heymann, B. and H. Grubmüller (2001). Molecular dynamics force probe simulations of antibody/antigen unbinding: entropic control and nonadditivity of unbinding forces. *Biophys J 81*(3), 1295–313.

Holmes, M., T. Buss, and J. Foote (1998). Conformational correction mechanisms aiding antigen recognition by a humanized antibody. *J Exp Med 187*(4), 479–85.

Howard, J. (1997). Molecular motors: structural adaptations to cellular functions. *Nature 389*(6651), 561–7.

Hsu, S., C. Peter, W. van Gunsteren, and A. Bonvin (2004). Entropy calculation of HIV-1 Env gp120, its receptor CD4 and their complex: an analysis of configurational entropy changes upon complexation. *Biophys J*.

Humphrey, W., A. Dalke, and K. Schulten (1996). Vmd: visual molecular dynamics. *J Mol Graph 14*, 33–38.

Izaguirre, J., D. Catarello, J. Wozniak, and R. Skeel (2001). Langevin stabilisation of molecular dynamics. *J Cem Phys 114*, 2090–2098.

Izrailev, S., S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten (1997). Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys J 72*, 1568–1581.

Jackson, R. and M. Sternberg (1995). A continuum model for protein-protein interactions: application to the docking problem. *J Mol Biol 250*(2), 258–75.

Janin, J. (1995). Elusive affinities. *Proteins 21*(1), 30–9.

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

Jones, S., A. Marin, and J. Thornton (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng 13*(2), 77–82.

Jones, S. and J. M. Thornton (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol 272*, 121–132.

Kaminski, G., H. Stern, B. Berne, R. Friesner, Y. Cao, R. Murphy, R. Zhou, and T. Halgren (2002). Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. *J Comput Chem 23*(16), 1515–31.

Karplus, M. and J. Kushick (1981). Method for estimating the conformational entropy of native proteins. *Macromolecules 14*, 325–332.

Keller, H. and P. Debrunner (1980). Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys Rev Lett 45*, 68–71.

Kellermayer, M. S. K., S. B. Smith, H. L. Granzier, and C. Bustamente (1997). Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science 276*, 1112.

Kendrew, J., R. Dickerson, B. Strandberg, R. Hart, D. Davis, D. Phillips, and V. Shore (1960). Structure of myoglobin. a three-dimensional fourier synthesis at 2 å resolution. *Nature 185*, 422–427.

Kimura, S., R. Brower, S. Vajda, and C. Camacho (2001). Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys J 80*(2), 635–42.

Kirschner, K., M. Eigen, R. Bittman, and B. Voigt (1966). The binding of nicotinamide adenine dinucleotide to yeast glyceraldehyde-3-phosphate dehydrogenase: Temperature-jump relaxation studies on the mechanism of an allosteric enzyme. *Proc Natl Acad Sci USA 56*, 1661–67.

Kitano, H. (2002). Computational systems biology. *Nature 420*(6912), 206–10.

Klimov, D. K. and D. Thirumalai (2000). Native topology determines force-induced unfolding pathways in globular proteins. *Proc Natl Acad Sci U S A 97*, 7254–7259.

Kollman, P., R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille (1997). *The development/application of a 'minimalist' organic/biochemical molecular mechanics force field using a combination of ab initio calculations and experimental data*, Volume 3 of *Computer Simulation of Biomolecular Systems*, pp. 83–96. Elsevier.

Koradi, R., M. Billeter, and K. Wuthrich (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph 14*(1), 51–5, 29–32.

Koshland, D. (1958). Application of a theory of enzyme specifity to protein synthesis. *Proc Natl Acad Sci USA 44*, 98–104.

Krammer, A., H. Lu, B. Isralewitz, K. Schulten, and V. V. (1999). Forced unfolding of the fibrnectin type III module reveals a tensile molecular recognition switch. *Proc Natl Acad Sci U S A 96*, 1351–1356.

Kraulis, P. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of proteins. *J Appl Crystallog 24*, 946–950.

Kumar, S., B. Ma, C. Tsai, N. Sinha, and R. Nussinov (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci 9*(1), 10–9.

Kuntz, I., J. Blaney, S. Oatley, R. Langridge, and T. Ferrin (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol 161*(2), 269–88.

Kyte, J. (2003). The basis of the hydrophobic effect. *Biophys Chem 100*, 193–203.

Lancet, D. and I. Pecht (1976). Kinetic evidence for hapten-induced conformational transition in immunoglobin MOPC 460. *Proc Natl Acad Sci U S A 73*(10), 3549–53.

Lazaridis, T. and M. Karplus (2000). Effective energy functions for protein structure prediction. *Curr Opin Struct Biol 10*(2), 139–45.

Leder, L., C. Berger, S. Bornhauser, H. Wendt, F. Ackermann, I. Jelesarov, and H. Bosshard (1995). Spectroscopic, calorimetric, and kinetic demonstration of conformational adaptation in peptide-antibody recognition. *Biochemistry 34*(50), 16509–18.

Lee, A., S. Kinnear, and A. Wand (2000). Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat Struct Biol 7*(1), 72–7.

Lee, J. C. and D. E. Discher (2001). Deformation-enhanced fluctuations in the red cell skeleton with theoretical relations to elasticity, connectivity, and spectrin unfolding. *Biophys J 81*, 3178–3192.

Lenne, P. F., A. J. Raae, S. M. Altmann, M. Saraste, and J. K. Horber (2000). States and transitions during forced unfolding of a single spectrin repeat. *FEBS Lett 476*, 124–128.

Levinthal, C. (1968). Are there pathways for protein folding? *J Chim Phys 65*, 44.

Li, H., A. Oberhauser, S. Fowler, J. Clarke, and J. Fernandez (2000). Atomic force microscopy reveals the mechanical design of a modular protein. *Proc Natl Acad Sci U S A 97*(12), 6527–6531.

Linke, W. and H. Granzier (1998). A spring tale: new facts on titin elasticity. *Biophys J 75*(6), 2613–4.

Lo Conte, L., C. Chothia, and J. Janin (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol 285*(5), 2177–98.

Lu, H., B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten (1998). Unfolding of titin immunogobulin domains by steered molecular dynamics simulation. *Biophys J 75*, 662–671.

Lu, H. and K. Schulten (1999). Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins 35*, 453–463.

Lu, H. and K. Schulten (2000). The key event in force-induced unfolding of titin's immunoglobulin domains. *Biophys J 79*, 51–65.

Luque, I. and E. Freire (2000). Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins Suppl 4*, 63–71.

Ma, B., T. Elkayam, H. Wolfson, and R. Nussinov (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A 100*, 5772–5777.

MacDonald, R. I. and E. V. Pozharski (2001). Free energies of urea and of thermal unfolding show that two tandem repeats of spectrin are thermodynamically more stable than a single repeat. *Biochemistry 40*, 3974–3984.

Marszalek, P. E., H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez (1999). Mechanical unfolding intermediates in titin modules. *Nature 402*, 100–103.

McAdams, H. and A. Arkin (1999). It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet 15*(2), 65–9.

McGough, A. (1999). How to build a molecular shock absorber. *Curr Biol 9*, R887–9.

Mehta, A., M. Rief, J. Spudich, D. Smith, and R. Simmons (1999). Single-molecule biomechanics with optical methods. *Science 283*(5408), 1689–95.

Mendez, R., R. Leplae, L. De Maria, and S. Wodak (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins 52*(1), 51–67.

Mercier, P., L. Spyracopoulos, and B. Sykes (2001). Structure, dynamics, and thermodynamics of the structural domain of troponin C in complex with the regulatory peptide 1-40 of troponin I. *Biochemistry 40*(34), 10063–77.

Miao, J., H. Chapman, J. Kirz, D. Sayre, and K. Hodgson (2004). Taking X-ray diffraction to the limit: macromolecular structures from femtosecond X-ray pulses and diffraction microscopy of cells with synchrotron radiation. *Annu Rev Biophys Biomol Struct 33*, 157–76.

Minajeva, A., M. Kulke, J. Fernandez, and W. Linke (2001). Unfolding of titin domains explains the viscoelastic behavior of skeletal myofibrils. *Biophys J 80*(3), 1442–51.

Monod, J., J. Wyman, and J. Changeux (1965). On the nature of allosteric transitions: A plausible model. *J Mol Biol 12*, 88–118.

Najmanovich, R., J. Kuttner, V. Sobolev, and M. Edelman (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins 39*(3), 261–8.

Neutze, R., R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu (2000). Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature 406*(6797), 752–7.

Northrup, S. and H. Erickson (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci U S A 89*(8), 3338–42.

Noskov, S. and C. Lim (2001). Free energy decomposition of protein-protein interactions. *Biophys J*, 737–750.

Oberhauser, A., P. Hansma, M. Carrion-Vazquez, and J. Fernandez (2001). Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc Natl Acad Sci U S A 98*(2), 468–472.

Oberhauser, A., P. Marszalek, H. Erickson, and F. J.M. (1998). The molecular elasticity of the extracellular matrix protein tenascin. *Nature 393*, 181–5.

Paci, E. and M. Karplus (1999). Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J Mol Biol 288*, 441–459.

Paci, E. and M. Karplus (2000). Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc Natl Acad Sci U S A 97*, 6521–6526.

Parak, F., E. Knapp, and D. Kucheida (1982). Protein dynamics. Mossbauer spectroscopy on deoxymyoglobin crystals. *J Mol Biol 161*(1), 177–94.

Pascual, J., J. Castresana, and M. Saraste (1997). Evolution of the spectrin repeat. *BioEssays 19*(9), 811–7.

Pascual, J., M. Pfuhl, D. Walther, M. Saraste, and M. Nilges (1997). Solution structure of the spectrin repeat: a left-handed antiparallel triple-helical coiled-coil. *J Mol Biol 273*, 740–751.

Perutz, M. (1970). Stereochemistry of cooperative effects in haemoglobin. *Nature 228*(5273), 726–39.

Perutz, M., A. Wilkinson, M. Paoli, and G. Dodson (1998). The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu Rev Biophys Biomol Struct 27*, 1–34.

Petsko, G. and D. Ringe (1984). Fluctuations in protein structure from X-ray diffraction. *Annu Rev Biophys Bioeng 13*, 331–71.

Rajamani, D., S. Thiel, S. Vajda, and C. Camacho (2004). Anchor residues in protein-protein interactions. *Proc Natl Acad Sci U S A 101*(31), 11287–92.

Raphael, R., A. Popel, and W. Brownell (2000). A membrane bending model of outer hair cell electromotility. *Biophys J 78*, 2844–2862.

Rief, M., J. Fernandez, and H. Gaub (1998). Elastically coupled two-level systems as a model for biopolymer extensibility. *Phys Rev Lett 81*(21), 4764–4767.

Rief, M., M. Gautel, F. Oesterhelt, F. Fernandez, and H. Gaub (1997). Reversible unfolding of indivdual titin immunoglobin domains by AFM. *Science 276*, 1109–1112.

Rief, M., J. Pascual, M. Saraste, and H. Gaub (1999). Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J Mol Biol 286*, 553–561.

Ritchie, D. (2003). Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins 52*(1), 98–106.

Ritchie, D. and G. Kemp (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*(2), 178–94.

Roux, B. and T. Simonson (1999). Implicit solvent models. *Biophys Chem 78*, 1–20.

Sarkar, A., R. Robertson, and J. Fernandez (2004). Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. *Proc Natl Acad Sci U S A 101*(35), 12882–6.

Sayle, R. A. and E. J. Milner-White (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci 20*, 374.

Schief, W. and J. Howard (2001). Conformational changes during kinesin motility. *Curr Opin Cell Biol 13*(1), 19–28.

Schlichting, I. and K. Chu (2000). Trapping intermediates in the crystal: ligand binding to myoglobin. *Curr Opin Struct Biol 10*(6), 744–52.

Schlitter, J. (1993). Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem Phys Lett 215*, 617–621.

Schneidman-Duhovny, D., Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H. Wolfson (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins 52*(1), 107–12.

Schrauber, H., F. Eisenhaber, and P. Argos (1993). Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol 230*(2), 592–612.

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

Schwarz, F., D. Tello, F. Goldbaum, R. Mariuzza, and R. Poljak (1995). Thermodynamics of antigen-antibody binding using specific anti-lysozyme antibodies. *Eur J Biochem 228*(2), 388–94.

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

Sharp, K., A. Nicholls, R. Fine, and B. Honig (1991). Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science 252*(5002), 106–9.

Simmerling, C., B. Strockbine, and A. Roitberg (2002). All-atom structure prediction and folding simulations of a stable protein. *J Am Chem Soc 124*(38), 11258–9.

Sleep, J., D. Wilson, R. Simmons, and W. Gratzer (1999). Elasticity of the red cell membrane and its relation to hemolytic dissorders: an optical tweezers study. *Biophys J 77*, 3085–3095.

Smith, S., L. Finzi, and C. Bustamente (1992). Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science 258*, 1122–1126.

Steinbach, P., R. Loncharich, and B. Brooks (1991). The effects of environment and hydration on protein dynamics: A simulation study of myoglobin. *Chem Phys 158*, 383–94.

Sundberg, E. and R. Mariuzza (2000). Luxury accommodations: the expanding role of structural plasticity in protein-protein interactions. *Structure Fold Des 8*(7), R137–42.

Sundberg, E., M. Urrutia, B. Braden, J. Isern, D. Tsuchiya, B. Fields, E. Malchiodi, J. Tormo, F. Schwarz, and R. Mariuzza (2000). Estimation of the hydrophobic effect in an antigen-antibody protein-protein interface. *Biochemistry 39*(50), 15375–87.

Teeter, M. and D. Case (1990). Harmonic and Quasiharmonic Descriptions of Crambin. *J Phys Chem 94*, 8091–8097.

Teusink, B., M. Walsh, K. van Dam, and H. Westerhoff (1998). The danger of metabolic pathways with turbo design. *Trends Biochem Sci 23*(5), 162–9.

Tidor, B. and M. Karplus (1994). The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J Mol Biol 238*(3), 405–14.

Tskhovrebova, L., J. Trinick, J. A. Sleep, and R. M. Simmons (1997). Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature 387*, 308–308.

Valencia, A. and F. Pazos (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol 12*(3), 368–73.

van Gunsteren, W. and H. Berendsen (1977). Algorithms for macromolecular dymaics and constraint dynamics. *Mol Phys 34*, 1311–27.

Viñals, J., A. Kolinski, and J. Skolnick (2002). Numerical Study of the Entropy Loss of Dimerization and the Folding Thermodynamics of the GCN4 Leucine Zipper. *Biophys J 83*, 2801–2811.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J Mol Graph 8*(1), 52–6, 29.

Wang, C., N. Pawley, and L. Nicholson (2001). The role of backbone motions in ligand binding to the c-Src SH3 domain. *J Mol Biol 313*(4), 873–87.

Wang, W., O. Donini, C. Reyes, and P. Kollman (2001). Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct 30*, 211–43.

Wasenius, V., M. Saraste, P. Salven, M. Eramaa, L. Holm, and V. Lehto (1989). Primary structure of the brain alpha-spectrin. *J Cell Biol 108*(1), 79–93.

Weiss, S. (1999). Fluorescence spectroscopy of single biomolecules. *Science 283*(5408), 1676–83.

Yao, H., D. Kristensen, I. Mihalek, M. Sowa, C. Shaw, M. Kimmel, L. Kavraki, and O. Lichtarge (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J Mol Biol 326*(1), 255–61.

Zaccai, G. (2000). How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science 288*(5471), 1604–7.

Zhang, C. and S. Kim (2003). Overview of structural genomics: from structure to function. *Curr Opin Chem Biol 7*(1), 28–32.

Zhong, R., D. Bourgeois, J. Helliwell, K. Moffat, V. Srajer, and B. Stoddard (1999). Laue crystallography: comming of age. *J Synchrotron Rad 6*, 891–917.

Zhou, H. (2001). Disparate ionic-strength dependencies of on and off rates in protein-protein association. *Biopolymers 59*(6), 427–33.

Zidek, L., M. V. Novotny, and M. J. Stone (1999). Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nat Struct Biol 6*, 1118–1121.