Protein Structure Prediction and Molecular Forces

Manfred J. Sippl

Contents

Abstract

Our ability to predict protein structures from amino acid sequences depends on our understanding of molecular forces. The same applies to the validation of protein structures determined in the laboratory. The protein structures available in the public domain contain a number of deficiencies and inconsistencies. As long as we are unable to recognize and correct such errors on a regular basis we cannot completely trust the experimental results and we cannot have any confidence in predicted structures. Here we investigate the theory of intermolecular forces from the perspective of protein structure theory, comment on the fundamental ideas involved, discuss the difficulties encountered, and we provide examples to illustrate the current state of affairs in protein structure validation and prediction.

1 Introduction

For more than half a century the protein folding problem challenges our understanding of physical systems on the molecular level. There seems to be a general consent regarding the view that the protein folding problem is at least qualitatively understood and that we do not need to invent new physical concepts or theories to explain the phenomenon. On the other hand we are still unable to predict structures with any confidence and we are still far from the precision obtained in experimental structure determination.

In an attempt to assess the current state of affairs in protein structure prediction an unbiased observer is bound to get confused. On the one hand he finds that the literature of the last fifty years abounds with claims that the protein folding problem has been solved. And on the other hand he finds that recently published protein structures determined with high precision are flawed with an astonishing amount of errors and deficiencies. Our observer might wonder how this fits together and what he can trust and what not.

To be specific consider Figure 1. Here we see a small portion of the high resolution crystal structure of dethiobiotin synthetase determined to 1.6 Å resolution. The amide nitrogen of asparagine 52, which is a hydrogen bond donor, is in hydrogen bond contact with two other nitrogen atoms, whereas the oxygen atom of the same asparagine, a hydrogen bond acceptor, is in hydrogen bond contact with two other hydrogen bond acceptors. But this cannot be true. The hydrogen bond donors carry positive partial charges, whereas the hydrogen bond acceptors carry negative partial charges and hence the configuration found in the crystal structure is strongly repulsive violating elementary physico-chemical standards.

Obviously, we can correct the problem by flipping the amide plane as shown in Figure 1 (b). Now there are four ideal hydrogen bonds between the amide group of asparagine 52 and the surrounding hydrogen bond donors and acceptors. The difference in energy between the correct and the incorrect conformation of figure 1 is enormous. In the former we have four strong hydrogen bonds, in the latter these are substituted by highly unfavorable interactions creating the antidote of four hydrogen bonds. In fact, the energy difference between incorrect and correct configuration is so high that the protein disintegrates when this energy is released at once.

The problem seems easy enough to suspect that in experimentally determined protein structures this kind of error is the exception rather than the rule. But as a matter of fact one out of five amide groups of asparagine and glutamine side chains is found in the incorrect conformation. As a consequence the 30,000 protein structures currently available in the public domain contain more than half a million erroneous rotamers of asparagine and glutamine side chain amides.

The problem originates from the fact that electron densities around the amide nitrogen and oxygen atoms of asparagine and glutamine residues are quite similar so that the locations of these atoms can be determined to high precision but not their identity. Hence, the problem seems to be specific for X-ray analysis of protein crystals. However, the error rate in structures determined by nuclear magnetic resonance (NMR) is even slightly higher. The experimental data obtained from X-ray and NMR experiments are generally refined by

various refinement protocols before the final coordinates are deposited and we have to conclude that the refinement protocols used are unable to correct unfavorable configurations or perhaps that they even introduce such errors.

Inspecting the example shown in figure 1 and similar cases (e.g. figures 2 and 3) and applying basic physico-chemical principles, we have not much difficulty in spotting the errors by eye, although there are other cases where this is not so obvious (e.g. see figure 1 of [49]). However, in presenting the case we have colored and oriented the atoms and labeled the atomic distances in ways to make the problem visible. In general such groups are hidden among thousands of atoms and they are not easy to find by visual inspection.

On the other hand it is generally assumed that the laws governing molecular interactions are known with sufficient quantitative precision so that such errors should be detectable by straightforward energy calculations. But again, this is more difficult than it sounds. First of all, there are many signs that tell us that the current models of molecular interactions are inadequate as we will discuss below. And second, X-ray analysis of protein crystal structures generally do not yield the positions of hydrogen atoms, but these atoms are required by most of the molecular force fields and energy functions currently in use. In particular, the energy of hydrogen bonds calculated by such molecular force fields strongly depends on the position of the proton which is shared between hydrogen bond donor and acceptor. Before any energy calculations can be done, the positions of the protons must be specified. The trouble is that the locations of the protons are not available from experiment and hence there is considerable freedom in the choice of proton positions. This choice may be guided by chemical intuition but some degree of arbitrariness necessarily remains. Hence, we are confronted with two difficult problems: the uncertainty in proton positions and the uncertainty in the quality of energy functions.

This introductory example shows that even our best structures contain an astonishing amount of errors. But this is only the tip of an iceberg. On a local scale there are similar problems with histidine rotamers, poorly resolved loops, disordered regions, etc. and on a global scale the public domain contains partially erroneous as well as completely misfolded structures. It is one of the major challenges of protein structure theory and the duty of curators of data bases to provide methods which can detect and correct such errors. Without convincing achievements in this direction attempts to predict protein structures remain elusive, we cannot have any confidence in predicted structures, and we even have to doubt the validity of experimental results.

Over the years there have been many successful predictions that seem to contradict such arguments and in particular the repeated CASP experiments have produced quite amazing predictions. But essentially all predictions reported in the literature rely heavily on sequence and structure data bases and expert knowledge and frequently a simple key word or sequence search yields an excellent prediction in the form of a related structure. Such results may turn out to be extremely useful in addressing biological questions, solving a related structure, studying biochemical mechanisms, and so on, but they leave the most important question unanswered, that is, is the structure correct?

The data mining aspects of protein structure prediction and related approaches are well

covered by the CASP experiments and extensive reviews are available in the recent literature . There is no need to repeat this here. Instead we take the opportunity to reflect on some fundamental problems in protein structure theory. Our main focus are the molecular interactions found in proteins and in particular the hydrogen bond, ubiquitous in biological macromolecules. A commented summary of the bibliography is found in the last section of this report.

2 Molecular Forces in Quantum Mechanics

The problem of molecular forces and the cohesion of matter has a long and interesting history [33]. Today the definitive language of molecules is quantum mechanics and hence any approach to molecular interactions necessarily has to start with Schrödinger's equation and the associated wave functions. We will not go into details here, but we need a brief review of what quantum mechanical calculations are able to do today. This also sheds some light on the origin of the force fields currently used in structure prediction, molecular modeling, and molecular dynamics simulations of proteins and in the refinement protocols employed in protein structure determination by X-ray and NMR methods.

The problem with quantum mechanics generally is that either we have studied the subject in some detail so that we are fluent in the language, know how to construct and solve the Schrödinger equation, and have the necessary experience to interpret the results. Assimilation of the postulates and mastering the subject in some detail certainly requires a couple of years. If we are interested in proteins we usually cannot afford the time necessary to go into such detail. Hence, basic expressions of quantum mechanics, like the Schrödinger equation, are not very informative since an expert knows them by heart whereas they are unintelligible to the novice. But they are a necessary starting point as can be judged by any textbook on physics and chemistry written after 1926.

The basic recipe of molecular quantum mechanics goes like this: Write down the Schrödinger equation for the molecular system of interest. The resulting equation is always 'exact' but since we cannot solve the equation directly we have to apply an arsenal of elegant and ingenious techniques to find approximate solutions. The calculations are considered to be successful if the results resemble experimental data. If the discrepancy between theory and experiment is too large, we can go back and try other tricks or simplify the system. There is however, a certain danger to get lost in this loop.

To be specific, the Schrödinger equation is written in the compact notation

$$H\Psi = E\Psi, \tag{1}$$

where H is the Hamiltonian, also called the energy operator, of the molecular system, Ψ are the solutions, called eigenfunctions of the operator or wave functions, and E stands for the associated energy eigenvalues.

Neglecting spin and relativistic effects the Hamilton operator of a molecule composed

of n electrons and N atomic nuclei is

$$H = -\frac{\hbar}{2m}\sum_{i}^{n}\nabla_i^2 + \sum_{i,I}^{n,N}\frac{-e^2 Z_I}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{R}_I|} + \frac{1}{2}\sum_{i\neq j}^{n}\frac{e^2}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{r}_j|} \quad (2)$$

$$-\sum_{I}^{N}\frac{\hbar}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{I\neq J}^{N}\frac{Z_I Z_J e^2}{|\mathbf{R}_I - \mathbf{R}_J|}, \quad (3)$$

(a detailed discussion of this operator and the meaning of the various symbols is found in [23]). The resulting Schrödinger equation $H\Psi = E\Psi$ can be solved exactly for one electron and one nucleus. Exactly means that the associated eigenfunctions Ψ are obtained as elementary functions that can be computed to any desired precision. In all other cases only approximate solutions of E and Ψ can be found. Here 'approximate' means two different things. On the one hand it means that the values of the solutions come out as a table of function arguments versus function values. And on the other hand this means that the respective function values are only approximations to the true but unknown values. The reliability of the results obtained is then usually judged by comparison with experimental data.

The solutions Ψ have many different aspects. In the case of molecules consisting of more than one atom the nuclei are generally fixed in space, which is usually called the Born-Oppenheimer approximation, so that only the electrons need to be considered. For a specific configuration of the nuclei the positions of the electrons are then obtained from the scalar product of the wave functions Ψ.

The most interesting solution is the so called ground state, characterized by the lowest energy E_0 and the associated wave function Ψ_0. In chemical applications it usually suffices to know the ground state. But how do we know the positions of the nuclei? If we know the structure from experiment (e.g. X-ray), then, in principle, we can use the theory to confirm the structure. But in actual fact the argument usually goes the other way round. Since the solution of the Schrödinger equation for several atoms is a formidable computational challenge the experimental structure is usually used to judge the quality of the calculations. Deviations in the order of a few percentages of bond distances and bond angles between experimental structure and computed result are generally rated as an excellent agreement between theory and experiment.

But if we need to determine the unknown structure of a molecule by quantum mechanical calculations the only possibility is the systematic variation of the positions of the nuclei. For each configuration of the nuclei we then must find the associated ground state wave function for the electrons and the associated ground state energy. To find the most stable molecular structure we then have to find the minimum among all these ground state energies as a function of the positions of the nuclei. With the advent of density functional theory, and quantum Monte Carlo methods such computations have become feasible for small molecular systems, where small means a few atoms. The remarkable successes of these approaches have led to widespread interest in density functional theory as the most promising approach for accurate, practical methods in the theory of materials [23].

We must add however, that in the actual calculation of molecular systems we still have to deal with a series of nasty problems. Except for the hydrogen atom, we have to replace the actual Hamiltonian by a so called mean field approximation, so that we can treat the particles as independent. The reason for this is that presently we cannot properly handle the pairwise interactions of a system of more than two particles in our calculations. A large part of current quantum mechanical research is concerned with approaches that can be used to circumvent these obstacles.

As Paul Dirac noted in 1929:

> The underlying physical laws necessary for a mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.

(quoted in [33]). Since then the computational side of quantum mechanics has made considerable progress, but with respect to larger molecules Dirac's assessment of 1929 is still up to date. There may be surprises ahead but so far the problems encountered in protein structure validation and prediction are not amenable to direct quantum mechanical computations. To make progress we have to try other approaches.

3 Molecular Mechanics Force Fields

Although direct quantum mechanical calculations are rarely applied to problems in protein structure theory some quantum mechanical results have found their way into protein structure prediction and refinement. The most popular model for pairwise atomic interactions is the Lennard-Jones potential

$$E(r) = -\frac{A}{r^6} + \frac{B}{r^{12}}. \tag{4}$$

This function describes the potential energy of the interaction of two noble gas atoms as a function of particle separation r. Most textbooks of biochemistry, physical chemistry, protein structure, and so on, explain the interactions found in proteins and other biological molecules in terms of this model and thus for many the Lennard-Jones potential has acquired the status of a physical law. To a large extent the Lennard-Jones potential owes its popularity to the simple functional form and the comprehensible physics embodied in this model. This is in stark contrast to the Copenhagen interpretation of quantum mechanical phenomena which according to Niels Bohr are necessarily incomprehensible to the human mind.

The attractive term, r^6, originates from the interaction of the two instantaneous dipole moments of the interacting atoms and the repulsive term, r^{12}, is due to the Coulomb repulsion of the core electrons. This energy function is a quantum mechanical result and its development has an interesting history [33]. By all standards the Lennard-Jones potential is a very good model for the interaction of two argon atoms. However, it turned out that for any other material than a noble gas this is a very poor approximation [10].

The Lennard-Jones potential is a major component of the standard empirical potential energy functions frequently employed in problems of protein structure research. The complete energy function is

$$\begin{aligned} E &= \sum K_b(b-b_0)^2 \\ &+ \sum K_\theta(\theta-\theta_0)^2 \\ &+ \sum K_\phi\left[1+\cos(n\phi-\delta)\right] \\ &+ \sum_{i<j} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \\ &+ \sum_{i<j} \frac{q_i q_j}{D_0 r_{ij}}. \end{aligned} \tag{5}$$

The first two terms model the lengths of covalent bonds and the angles between two consecutive covalent bonds, respectively. The third term describes the variation of energy as a function of the angle of a rotatable bond. The fourth term is the Lennard-Jones potential for pairs of atoms and the last term is the electrostatic interaction of atom pairs modeled by the Coulomb interaction, where D_0 is the dielectric constant.

It is generally assumed that this 'minimalist model' [8] captures the main components of the conformational energy of protein molecules, in the sense that it seems to be a satisfactory compromise between simplicity and accuracy [3]. In fact the model captures the local stereochemistry of protein structures and it is a very useful tool to remove steric clashes and local distortions in structural models. However, there are a number of problems. The parameters describing the specific properties of the various atom types are estimated from a multitude of experimental data and quantum mechanical considerations. These parameters are interdependent and errors in one parameter are propagated and perturb the assessment of other parameters (see e.g. [44]). But the most serious problem is the assumption that all non-bonded interactions can be modeled by a combination of a Lennard-Jones type potential and a Coulomb interaction. It has never been demonstrated that this assumption is justified.

To the contrary, Novotny, Bruccoleri and Karplus [30] have shown in 1984 that native and completely misfolded structures of proteins cannot be distinguished by energy calculations based on the model defined by equation 5. Subsequent studies reported improved performance when the aqueous environment is included in the calculations. Including water molecules in such calculations dramatically increases the complexity and computational demands. Among all problems involving molecular forces the problem of liquid water and its interactions with solutes stands out as the most challenging [7, 33] and Lennard-Jones potentials and Coulomb's law are inadequate models for these interactions. Adding water molecules in this scheme does not solve the problem.

4 Molecular Forces in Statistical Mechanics

Statistical mechanics provides the tools to investigate the configurations of a large number of particles and the associated molecular forces. One of the key results of classical statistical mechanics is Boltzmann's principle

$$p(x) = \frac{1}{Z} e^{-\epsilon(x)/kT}, \tag{6}$$

where

$$Z = \sum_x e^{-\epsilon(x)/kT}, \tag{7}$$

is called the Boltzmann sum or partition function. The principle states that the occupancy $p(x)$ of a certain state x of a system is a direct consequence of the energy $\epsilon(x)$ of this state. If x is the distance between two atoms of types a and b then $\epsilon_{a,b}(x)$ is the energy of interaction of these atoms at this distance and $p(x)$ is the associated probability density distribution that can be calculated from equation 6. The terms occupancy, density distribution and relative frequency are often used as synonyms for $p(x)$.

There are several routes that lead to Boltzmann's distribution. The approach frequently found in textbooks of statistical mechanics (e.g. [13]) is based on maximizing the number of microstates

$$\Omega = \frac{N!}{n_1! n_2! \cdots n_M!}, \qquad N = \sum_i^M n_i. \tag{8}$$

Here N is usually interpreted as the number of particles and M as the number of distinguishable states of the particles. We can be more general and interpret N as the total number of cases and M as the number of distinguishable cases. If we are interested in distances between atoms, for example, then N may represent the number of distinct pairs of atom types (a, b) found in our system and M may refer to the number of distinct distance intervals used to record the separation of two atoms. Whatever states the indices i may specify, the n_i represent the number of cases that are found in state i and we can express these occupancies also in terms of probabilities $p_i = n_i/N$. If our variables are, like distances, continuous we may use the indices to represent intervals. In this case we may write $p(x) = n(x)/n$. Note that there is no particular significance in using the symbols x, i, etc., since the index sets represented by these symbols are merely used to label the various states. A microstate is then defined by the exact specification of the particular state of each individual particle or case, whereas a macrostate is defined by the occupancies or frequencies n_i or, equivalently, by the probabilities p_i, where $i = 1, \ldots, M$. In general, a particular macrostate can be realized by many distinct microstates.

A central assumption (often called postulate) of statistical mechanics is that the equilibrium state of a system corresponds to that particular macrostate which has the maximum number of microstates, Ω, given the constraint of a constant number of cases,

$\sum n_i = N =$ constant, and given the constraint of constant energy U, i.e.

$$U = \sum_{i=1}^{M} n_i \epsilon_i = \text{constant} \tag{9}$$

The extremum of Ω is found by maximizing the logarithm of the number of microstates $\log \Omega$ by a variation of the absolute occupancies n_i under the given constraints. This immediately yields Boltzmann's distribution (equation 6), where the temperature T is obtained from the additional assumption (i.e. postulate) that the entropy S is proportional to the logarithm of the number of microstates, $S = k \log \Omega$ (k Boltzmann's constant), and the thermodynamic relationship $dS = dQ/T$, where dQ is the heat input or output of a reversible thermodynamic process at temperature T.

To apply Boltzmann's principle we need to specify the energy function ϵ_i or $\epsilon(x)$ governing the system of interest. From the energy function and Boltzmann's principle we can then compute the occupancies of the various distinguishable states n_i and the associated probabilities p_i.

Another route to the Boltzmann distribution (equation 6) is due to Williard Gibbs [12]. He starts out from the classical laws of motion of a system of particles. The motion is governed by the total energy, i.e. the sum of the kinetic and potential energy of the particles, which is called the Hamiltonian of the system. As it turns out there are some integrals of the motion that remain invariant if the system is in equilibrium. In a less technical language this means that some quantities remain constant even if the atoms or particles of the system are moving relative to each other. Two such quantities are the total energy and the probability density distribution expressed as a function of coordinates and momenta. This statement can be summarized in the following way

$$\sum_i \left(\frac{dP}{dq_i} \frac{d\epsilon}{dp_i} - \frac{dP}{dp_i} \frac{d\epsilon}{dq_i} \right) = 0, \tag{10}$$

where we have used the original notation of Williard Gibbs [12]. Here the probability density function $P = P(q_1, \ldots, p_1, \ldots)$ is a function of the coordinates q_i and momenta p_i and $\epsilon = \epsilon(q_1, \ldots, p_1, \ldots)$ is the energy of the system, which is again a function of coordinates and momenta. The sum extends over all particles $i = 1, \ldots N$. Gibbs then argues that a function of the form

$$P = e^{\frac{\Psi - \epsilon}{\Theta}} \tag{11}$$

seems to represent the most simple probability density function possible that satisfies equation 10. Here Θ, called the modulus of the distribution, is later shown to be equivalent to the temperature, whereas Ψ is a normalizing factor, defined by

$$e^{-\frac{\Psi}{\Theta}} = \int \cdots \int e^{-\frac{\epsilon}{\Theta}} dp_1 \cdots dq_n. \tag{12}$$

The integral is called partition function or Boltzmann sum which we have already encountered in its discrete form in equation 7.

Since the kinetic energy is a function of the momenta p_i whereas the potential energy is a function of the coordinates q_i it so happens that the partition function 'partitions' additively into a kinetic and a configurational part. This simplifies the calculations to some extent, since the equilibrium structure of a molecular system only depends on those variables needed to specify the structure of the system, like coordinates or distances between particles, while we may neglect the associated momenta.

The approach of Gibbs has the advantage that it is rooted in the most general version of classical mechanics, i.e. on the laws of motion expressed in Hamiltonian form. The main point is that the energy function of a system (i.e. the Hamiltonian) uniquely determines the probability density function $p(x)$. If the system is not disturbed by external forces, so that it remains in equilibrium, then the density distribution $p(x)$ is constant in time. To apply the theory we have to find or guess the energy function of the system from which we can deduce the density of states $p(x)$ and for this it does not matter whether we take the route of Boltzmann or that of Gibbs. But, as long as we do not know the functional form of the molecular interactions, $\epsilon(x)$, we cannot compute $p(x)$ and with respect to protein structure validation and prediction we are still at square one.

5 Molecular Forces and Radial Distribution Functions

A direct offspring of the principles of statistical mechanics is the theory of radial distribution functions and correlation functions which is the main subject of the theory of simple liquids [5,14,16]. This branch of statistical mechanics provides powerful tools for the investigation of molecular forces. The principal ideas of this theory are straightforward. Consider a system of particles or atoms. We may tag one of the atoms and determine the distribution of all other atoms as a function of the distance from the tagged atom. We may then repeat this analysis for all the remaining atoms and compute the average distribution. The result is a radial distribution function representing the average configuration of the particles in terms of inter-particle distances.

Radial distribution functions of liquid and solid materials can be determined by X-ray or neutron diffraction [5, 14]. These distribution functions can then be compared to theoretical results obtained from molecular simulations and other calculations. The essential relationship between molecular forces, interaction energies, and distribution functions is summarized in the following expression,

$$\overline{F}(r) = -\nabla \overline{U}(r) = kT \nabla \log[g(r)]. \tag{13}$$

Here r is the distance between two particles and $F(r)$ is the force between two particles. The force is the negative gradient, denoted by ∇, of the potential energy $U(r)$ between the two particles, and $g(r)$ is the radial distribution function. The bars over F and U remind us that these quantities are averages over all pairs of particles in the system. The symbol g denoting the radial distribution function has no bar since it is explicitly defined as an average over many particles.

The detailed expression for the radial distribution function has the form of Boltzmann's law (equation 6)

$$g(r_{1,2}) = \frac{1}{Z} \int \cdots \int e^{-U(\mathbf{r})/kT} dr_3 \cdots dr_N \tag{14}$$

where

$$Z = \int \cdots \int e^{-U(\mathbf{r})/kT} d\mathbf{r}. \tag{15}$$

Here $\mathbf{r}$ denotes the positions of all particles $r_1, r_2, \ldots, r_N$. To obtain $g(r_{1,2})$ we integrate the Boltzmann factor, $\exp[-U(\mathbf{r})/kT]$, over all particles except two and normalize by the so called configuration integral, i.e. the configurational part of the partition function, which is just the integral over all possible positions of the particles. These expressions may appear somewhat unwieldy and as a matter of fact they are. A major reason for this is that in the derivation of these expressions we have to start from a notation that allows us to handle the positions of all particles although the final result depends on the distance between two particles.

We can summarize and interpret all this in a quite simple form. The radial distribution function can be written as

$$g(r) = \frac{n(r)}{n_g(r)}, \tag{16}$$

where $n(r)$ is the number density, i.e. the number of particles, found at distance r from a central particle as observed in a material sample and $n_g(r)$ is the distribution of particles in an ideal gas having the same particle density as the sample. By definition, the particles in an ideal gas do not interact at all, i.e. there are no forces and the energy is the same for all distances, but the distribution contains all geometrical constraints which are impressed on the distribution of distances in a three dimensional space. Therefore, the ideal gas distribution $n_g(r)$ is a convenient reference state and the fraction, equation 16, measures the deviation of the distribution of distances in the material sample from the distribution of an ideal gas. For $g(r) > 1$ the interaction between two particles at distance r is attractive and it is repulsive for $g(r) < 1$.

From the radial distribution function we obtain an energy function

$$w(r) = -kT \log[g(r)], \tag{17}$$

describing the average energy of interaction of two particles in the sample. Since this function is derived from the averaged force $\overline{F}(r)$ acting between all pairs of particles (equation 13) energy functions of this type are called potentials of average force and also potentials of mean force.

Radial distribution functions have been determined for a number of 'simple liquids', including water. Of course, water is not a simple liquid. The term 'simple' refers to the number of distinct atom types constituting the liquid and not to the complexity of the interactions involved. The functions obtained from diffraction experiments can then be compared to radial distribution functions obtained from molecular simulations based on Boltzmann's distribution (equations 6 and 16). There is an extensive literature on simulations of systems

of hard spheres and systems of particles that interact according to Lennard-Jones type potentials [5, 14]. In general the basic shapes of the functions derived from such simulations agree at least qualitatively with the radial distribution functions obtained from experiment.

In general, radial distribution functions can be obtained only for simple molecular systems. The reason is that radial distribution functions measured on liquid or solid materials consisting of more than one atom type are combinations or superpositions of distinct atom pair radial distribution functions. For example, the radial distribution function of water is a superposition of three types of functions corresponding to the interactions of a pair of hydrogen atoms, a pair of oxygen atoms, and a pair of one hydrogen and one oxygen atom and it is not an easy matter to extract these individual components from the total radial distribution function.

Radial distribution functions and potentials of mean force emphasize the direct link between molecular forces and molecular structure. Of course, this connection is already implicitly contained in Boltzmann's law, equation 6, but the theory of radial distribution functions provides the appropriate mathematical tools to accomplish the transformations between energy and structure and it yields a plausible interpretation of the physical principles involved. The connection between energy and structure provides the means for the quantitative analysis of the forces and interactions in proteins as we will see shortly. Before we embark on this endeavor it is instructive to briefly contemplate the conditions under which physics was practiced when the foundations of statistical mechanics and quantum mechanics were framed and elaborated.

6 Molecular Forces and Physical Theories

The fundamental physical theories of thermodynamics, statistical mechanics, quantum mechanics, relativity, etc., and the associated mathematical tools have been invented within a period of a hundred years, which extends roughly from 1850-1950. The physicists of these days generally had a small number of fundamental experimental data and the physical theories were invented to explain the experimental facts. If this failed, attempts were made to find new basic principles. The most dramatic example is certainly the development of quantum mechanics. Starting from a few basic assumptions the theory developed between 1900 and 1930 eventually managed to comprehend and 'explain' a large body of important experimental facts, like black body radiation, atomic spectra, chemical bonding, etc., with unprecedented precision.

Starting with the second half of the twentieth century the data processing power of electronic computers opened possibilities that were unthinkable in the days of Boltzmann, Schrödinger, or Dirac. Detailed calculations involving a fistful of particles are impractical without a computer and limited capabilities in data processing and numerical work necessarily constrain the possible approaches to a physical problem. These limitations shaped the very definition of physics and what was considered a tractable physical problem. Even in the opinion of Boltzmann, proponent of the atomic theory of matter and inventor of the statistical theory of gases, the computation of distances and forces for a large number of

individual atoms was not an attractive option. In 1895 Boltzmann remarks,

> the difficulty of enumerating all the material points of the universe and of determining the law of mutual force for each pair, would only be a quantitative one; nature would be a difficult problem, but not a mystery for the human mind.

(quoted in [33]). The meaning of Boltzmann's prose is not entirely clear but he seems to say that a detailed description of a large number of specific molecular configurations and interactions is a boring venture and does not yield to general physical laws. We must concede to Boltzmann that he and his contemporaries never had the chance to see the atomic structure of a protein and in his time the existence of such complex but nonetheless well defined molecular structures must have been unthinkable. Today we have to face this complexity and to make progress we have to determine the laws of mutual force of all the various types of atomic interactions found in biological structures, may their number be large or small.

For this venture the methods of statistical mechanics, laid out by Boltzmann, Gibbs and many others, provide an appropriate basis. To proceed we emphasize one additional point regarding Boltzmann's law, equation 6. As it stands, the expression implies that the probability density, $p(x)$, is a function of the energy $\epsilon(x)$. Once we know the energy function $\epsilon(x)$, we are able to compute $p(x)$ and hence the structure of the system. And from the associated partition function, equation 7, we can derive thermodynamic quantities like entropy, Helmholtz free energy, and so on. An approach like this is a characteristic example of physical reasoning. In the language of physics, a true understanding and explanation of a natural phenomenon is achieved, if we can derive the energy function or the energy operator from first principles and if we can show that the laws of quantum mechanics, classical mechanics, and statistical mechanics when applied to this energy function reproduce the experimental facts to some degree of accuracy.

With this in mind we understand Boltzmann's reasoning that a proper description of atomic interactions requires that we can reduce the problem to a small set of basic laws of force. A large number of distinct energy functions, that need to be adapted to each individual case, is of little value. We also see that the energy operator, equation 3, used to express the physics of a molecular system in quantum mechanical terms, contains a simple electrostatic energy function and thus satisfies Boltzmann's demand. Neglecting spin, all the interactions among electrons and nuclei are governed by Coulomb's law, where the energy function has the most simple form $q_a q_b/r$, where the q's correspond to the charges carried by electrons and nuclei (i.e. the e and Z in equation 3), and where r is the separation of the particles. All the miracles of chemistry and biology seem to follow from this disarmingly simple function. But as we have already seen, the structures of proteins are beyond the present capabilities of this astonishing theory.

7 Molecular Forces and Molecular Structure

The energy function plays a central role in the description of a physical system and most other quantities, in particular density distribution functions, are considered to be quantities

of less nobility. Inverting Boltzmann's law (equation 6),

$$\epsilon(x) = -kT \log p(x) - kT \log Z \tag{18}$$

the energy $\epsilon(x)$ is obtained from the density distribution $p(x)$. When applied to molecular interactions the expression shows that the 'laws of force' can be derived from an experimental determination of the density distribution $p(x)$ of the configurational variables x, where the experiments are carried out at a certain temperature T. Here the energy has lost its dominant position and the two versions of Boltzmann's law show that in actual fact energy and density are on an equal footing and they are two sides of the same coin.

To apply the inverted form of Boltzmann's law in the analysis of atomic interactions we need a proper amount of highly resolved and reliable experimental data on the relative arrangements of atoms so that we can compile density distributions of sufficient accuracy. The amount of detailed structural information on proteins and on other molecular systems acquired over the last decades is enormous and suffices to compile potentials of average force for most interactions encountered in the structures of proteins.

We briefly review the steps required to compile potentials of mean force from a set of protein structures [37, 38]. We write Boltzmann's law in the form

$$E(a, b, r) = -kT \log g(a, b, r) \tag{19}$$

where $E(a, b, r)$ is the generic potential of mean force for the interaction of two atoms of type a and b at separation r, $g(a, b, r)$ is the generic two particle radial distribution function, and k is Boltzmann's constant and T is the absolute temperature. The potential function $E(a, b, r)$ may be split into an annealed and a quenched part. The annealed or 'soft' part is specific for the particular atom types a and b. The quenched or 'hard' contributions are intrinsic to the covalent structure and compact nature of protein structures and they also contain the subtle constraints imposed by the geometry of three dimensional space. The quenched contributions are extracted from the generic radial distribution functions by averaging over all atom pairs a and b which yields the unspecific radial distribution function

$$g(r) = \frac{1}{n} \sum_{a,b} g(a, b, r) \tag{20}$$

where n is the number of distinct generic distribution functions. The specific radial distribution functions are defined as the fractions

$$g_s(a, b, r) = \frac{g(a, b, r)}{g(r)}. \tag{21}$$

Here, in analogy to the ideal gas distribution of equation 16, the unspecific radial distribution function acts as a reference state. From the specific radial distribution functions $g_s(a, b, r)$ we obtain the specific potential of mean force

$$\epsilon(a, b, r) = -kT \log \frac{g(a, b, r)}{g(r)} = -kT \log g(a, b, r) + kT \log g(r). \tag{22}$$

In all these expressions we have implicitly used the convenient choice $-kT \log Z = 0$, but we could have used any other finite value since with respect to the potential of mean force at constant temperature T this term is an additive constant.

Equation 22 summarizes the basic recipe for the compilation of a set of potentials of mean force. The compilation of radial distribution functions and the application of equation 22 to the atom types found in protein structures requires a number of finer details that need to be considered. For example, some atom types are rare so that radial distribution functions compiled for these atom types may have rather large fluctuations as compared to the more frequent atom types. To handle this and other problems proper techniques are available [37].

8 Molecular Forces and Protein Structures

Work on mean force potentials for the interactions found in proteins has started in 1990 [37] and it was quickly demonstrated that these functions can be used to distinguish native from misfolded structures [15, 37] and several erroneous experimental structures were detected in the public domain [39, 40]. Concurrently the efficiency of mean force calculations enabled the development of new prediction techniques like fold recognition and threading [18, 21, 42, 46], and techniques like fragment assembly were used to correctly predict approximate structures of small proteins ahead of experiment [43]. In the following years mean force potentials were used to calculate the changes in stability associated with amino acid mutations [40], and the were used in protein design [51] and the docking and binding of small molecules to proteins [29]. An example of applications to protein sequence randomization and structural stability is summarized in figure 7. Early implementations of mean force potentials use simplified representations of protein structures considering only the C^{α} and C^{β} atoms along the polypeptide chain. One advantage of such reduced models is their computational efficiency enabling the search of large sequence and structure data bases and the generation and evaluation of a large number of conformations to find structures that are as close as possible to the native fold of a protein.

It is clear however, that the accurate validation and reliable correction of protein structures and the prediction of structures with a precision comparable to experimental structure determination requires force fields containing all the interactions found in proteins. The development of a complete set of mean force potentials [35, 41, 45] is the subject of current research [49, 50]. To illustrate the current state of the art we return to our introductory example of asparagine and glutamine rotamers.

The amide groups of asparagine and glutamine are specific examples of functional groups which act simultaneously as hydrogen bond donors and acceptors. Hydrogen bonds form when a strongly electronegative nitrogen or oxygen atom, the hydrogen bond donor, shares a covalently attached proton with a lone pair of electrons of another oxygen atom, called the hydrogen bond acceptor. In proteins hydrogen bond donors and acceptors are always an integral part of the polypeptide backbone or the amino acid side chains. Since molecular forces strongly depend on the covalent structure and chemical environment in the vicinity of the interacting atoms there are many individual types of hydrogen bonds in

proteins.

The variability of hydrogen bonds and other interactions is easily captured by mean force potentials. The only requirement is that we distinguish among the various atom types and compile potentials for the individual interactions. In particular the backbone atoms N, C^{α}, C', and O of the various amino acids are distinct atom types. With this distinction the standard amino acid residues found in protein crystal structures determined by X-ray analysis contain $n = 167$ distinct atom types resulting in a total of $(n+1) \times n/2 = 14,028$ interactions. This number does not include any explicit interactions involving hydrogen atoms, atoms of non-standard groups, or water. But these interactions are not completely neglected. Mean force potentials capture contributions from the complete chemical environment even if the latter is not completely specified. Therefore, an immediate advantage of potentials of mean force is that the evaluation of interaction energies between hydrogen bond donors and acceptors does not require the explicit consideration of hydrogen atoms.

However, it is known in advance that the data base contains a substantial fraction of incorrect asparagine and glutamine rotamers and at the outset it is unclear to what extent the error rate of more than 20% corrupts the radial distribution functions resulting in defective or unusable mean force potentials. The compilation of mean force potentials from defective data is in fact an intriguing challenge.

As it turns out the effect of errors on the potentials is quite small. Moreover, a refinement cycle consisting of rotamer correction and recompilation of potentials quickly converges to a stable solution and there is excellent agreement between the rotamer flips suggested by a thorough analysis based on physico-chemical principles [52] and those suggested by mean force calculations. Figures 1, 2, and 3 provide examples of incorrect rotamers as found in highly resolved protein crystal structures and the respective corrected configurations. Examples of potentials before and after refinement are shown in figure 4.

Potentials of mean force have several remarkable properties and difficult problems of protein structure analysis and prediction can be tackled quite successfully. The reason for this is that mean force potentials provide a concise and compact representation of the complete experimental knowledge on protein structures currently available. The only assumption or 'postulate' is that the relationship between structure and energy, as expressed by Boltzmann's law (equation 22) holds true. There are no additional physical parameters that enter the calculations, like the A's and B's of Lennard-Jones potentials, dielectric constants, partial charges, and so on, required to implement a minimal version of molecular force fields. Similarly, energy calculations based on potentials of mean force are straightforward and efficient and there is no need for sophisticated mathematical procedures as they are required for the solution of quantum mechanical differential equations. A most attractive property of mean force potentials is that they are self-consistent and self-correcting. Errors and inconsistencies in the data from which mean force potentials are compiled are detected by the very same potentials [50].

9 Molecular Forces of Hydrogen Bonds

Figure 5 shows the 'law of force' for four types of interactions found in protein structures. Each interaction is represented by two functions. The radial distribution functions are compiled from a data base of 833 highly resolved protein crystal structures. The solid lines correspond to potentials of mean force obtained from single protein chains and the respective radial distribution functions solely contain intramolecular distances. The dashed lines correspond to mean force potentials compiled from complete protein crystal structures. Here the radial distribution functions contain inter- as well as intramolecular distances. Nevertheless the two types of functions are quite similar demonstrating that intra- and intermolecular interactions follow the same basic rules and it also shows that the functional form of the individual potentials are quite stable with respect to variations in the data sources. Similar results are obtained when data sets are split into independent subsets. A comparison of independent data sets generally reveals that corresponding potentials are practically indistinguishable.

The interactions found in proteins are often grouped in polar, ionic, hydrophobic and hydrogen bond interactions. Figures 4 and 5 show a hydrogen bond interaction (a), a polar interaction (b), a hydrophobic interaction (c), and an interaction between hydrophobic and a polar groups (d). The interaction between the C^{γ} group of valine and the C^{δ} group of leucine (c), a typical hydrophobic interaction, has a deep minimum and is attractive over the whole distance range except for short distances where the atoms start to penetrate each other. The interaction between the hydrophobic C^{β} atom of valine and the the charged O^{δ} atom of aspartic acid (d) is predominantly repulsive. The interaction between the asparagine N^{δ_2} atom and the glycine backbone nitrogen (b) has an appreciable attractive energy well. This potential describes the interaction of two positive partial charges and one might expect that the interaction should be repulsive. However, the minimum is at the rather large distance of 5 Å. The attraction of these two atoms is mediated by a water molecule that is sandwiched between the two nitrogen atoms or some other group that can compensate the partial charges on the nitrogen atoms. This example demonstrates that the potentials encode the locations of intervening solvent molecules although the positions of these molecules are not explicitly specified.

The interaction between the N^{δ_2} atom of asparagine and the backbone oxygen of glycine (a) is a typical example of a hydrogen bond interaction. The potential has a deep narrow minimum at hydrogen bond contact corresponding to a short distance of 2.7 Å, separated by a high energy barrier from larger distances. Quite generally, mean force hydrogen bond interactions have the form of a molecular lock or shutter, characterized by an energy barrier with a maximum at r_m, separating a narrow energy well at hydrogen bond contact at r_c from large distances, as shown in figure 6.

The molecular lock is a general model for spatially precise and kinetically stable interactions where precision is determined by the width of the energy well and stability is mediated by the hight of the energy barrier, $\epsilon(r_m) - \epsilon(r_c)$. A consequence of this model is that stable bonds can be formed even if the energy balance of bond formation is positive,

zero, or only slightly negative and all these types of hydrogen bonds are actually observed in protein structures [50]. The barrier is a specific feature of hydrogen bonds and highly polar interactions. Other interactions, in particular hydrophobic 'bonds', have comparatively broad minima and they lack barriers. Such interactions are less precise and their stability is determined by ϵ_f the depth of the energy minimum as shown in figure 6 (d).

Proteins have a large number of stabilizing hydrogen bond interactions although the free energy differences between folded and unfolded states is comparatively small. A generally accepted qualitative explanation for this apparent paradox is that in the folding of proteins intermolecular hydrogen bonds between protein atoms and water molecules are replaced by intramolecular hydrogen bonds between protein atoms on the one hand and hydrogen bonds between water molecules on the other, where the liberation of water molecules from the protein compensates for the loss in entropy of the 'freezing' protein. The individual contributions of the various energy terms are large but they almost cancel so that the resulting total energy difference between folded and unfolded proteins is small. Failures to compute the energy difference of protein folding from theoretical models are generally attributed to uncertainties arising from the cancelation of large energy terms.

In contrast, a resolution of this paradox in terms of mean force potentials is straightforward. Hydrogen bonds are molecular locks that stabilize protein structures whether the free energy balance of bond formation is negative, positive, or zero. Even if the total sum of all interactions is comparatively small the structure of a protein can be quite stable since the energy barriers prevent unfolding. Nevertheless, a final resolution of such paradoxes and other miracles of protein folding requires that we are able to compute protein structures with a precision and confidence that rivals experimental structure determination and that we can reproduce the free energies, heat capacities, etc., obtained in the laboratory. After all, this is still a considerable challenge.

10 Bibliographical Notes

The amount of available literature on molecular forces, the theory of protein structure, and protein structure prediction is enormous. Our aim here is to provide a set of general pointers to the literature and to extend the references mentioned in the text.

The history of molecular forces from Newton to the present day is laid out in a recent book by Rowlinson [33]. There are many excellent texts on quantum mechanics. The basic concepts discussed here can be found in the recent monograph of Martin [23] (where equation 3 is discussed in some detail) or Atkins and Friedman [1]. Both books provide the technical details required for quantum chemical computations. The books by Cramer [6] and Thijssen [48] are specifically dedicated to computational work in quantum mechanics. A recent text on the quantum mechanics of intermolecular forces is the book by Finnis [10].

Empirical force fields and their applications in molecular simulations are extensively covered by Leach [20] and others [3]. A recent update of force field parameters is found in [19]. The history of the Lennard-Jones potential is extensively covered by Rowlinson [33] and in a more technical form by Finnis [10]. There are numerous text books covering

classical statistical mechanics. Brief accounts of the basic principles are found in [5, 13] and it is certainly advisable to consult Gibbs' elementary principles of statistical mechanics [12]. An accessible account of the theory of radial distribution functions and liquids is provided by Chandlers book [5] whereas the monograph by Hansen and McDonald [14] is a detailed account of the subject. A classic text in this area is the book by Hill [16]. The statistical mechanics of chain molecules, highly relevant to protein structure theory, is now out of fashion. The books by Flory [11] and Rubinstein and Colby [34] are excellent introductions.

Mean force potentials for proteins are introduced in [37] and reviewed in [38,40]. A recent publication emphasizing the informational theoretical aspects is provided by Solis and Rackovsky [47]. Examples of various applications and extensions of mean force potentials are found in [4, 22, 24, 25, 32, 36, 53].

An introduction to the main aspects of protein structure theory and protein structure prediction is found in the book of Finkelstein and Ptitsyn [9]. A recent review on fold recognition and related techniques is provided by Mizuguchi [26]. Examples of recent reviews on various aspects of protein structure prediction are [2, 17, 27, 31] and the entry point to the most recent CASP experiment is [28].

References

[1] P. W. Atkins and R. S. Friedman. *Molecular Quantum Mechanics.* Oxford University Press, 3 edition, 1997.

[2] Philip Bradley, Kira M S Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, Sep 2005.

[3] Charles Brooks, Martin Karplus, and Montgomery Pettitt. *Proteins. A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. John Wiley & Sons, 1988.

[4] Nicolae-Viorel Buchete, John E Straub, and Devarajan Thirumalai. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci*, 13(4):862–874, Apr 2004.

[5] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987.

[6] Christopher J. Cramer. *Essentials of Computational Chemistry.* Wiley, 2002.

[7] David Eisenberg and Walter Kauzmann. *The Structure and Properties of Water.* Oxford University Press, 1969.

[8] Alan Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W H Freeman, 1999.

[9] Alexei V. Finkelstein and Oleg B. Ptitsyn. *Protein Physics*. Academic Press, 2002.

[10] Mike Finnis. *Interatomic Forces in Condensed Matter*. Oxford University Press, 2004.

[11] Paul J. Flory. *Statistical Mechanics of Chain Molecules.* Wiley, 1969.

[12] J. Williard Gibbs. *Elementary Principles of Statistical Mechanics.* Ox Bow Press, 1901.

[13] A. M. Glazer and J. S. Wark. *Statistical Mechanics. A Survival Guide.* Oxford University Press, 2001.

[14] Jean-Pierre Hansen and Ian R. McDonald. *Theory of Simple Liquids.* Elsevier, 2006.

[15] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J Mol Biol*, 216(1):167–180, Nov 1990.

[16] Terrell L. Hill. *Statistical Mechanics. Principles and Selected Applications.* McGraw-Hill, 1956.

[17] Jol Janin and Michael Levitt. Theory and simulation Accuracy and reliability in modelling proteins and complexes. *Curr Opin Struct Biol*, 16(2):139–141, Apr 2006.

[18] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, Jul 1992.

[19] Elmar Krieger, Tom Darden, Sander B Nabuurs, Alexei Finkelstein, and Gert Vriend. Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins*, 57(4):678–683, Dec 2004.

[20] Andrew R. Leach. *Molecular Modelling. Principles and Applications.* Prentice Hall, 2 edition, 2001.

[21] R. Lthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85, Mar 1992.

[22] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–232, Aug 2001.

[23] Richard M. Martin. *Electronic Structure*. Cambridge University Press, 2004.

[24] F. Melo, R. Sanchez, and A. Sali. Statistical potentials for fold assessment. *Protein Sci*, 11(2):430–448, 2002.

[25] Sanzo Miyazawa and Robert L Jernigan. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J Chem Phys*, 122(2):024901, Jan 2005.

[26] Kenji Mizuguchi. Fold recognition for drug discovery. *Drug Discovery Today*, 3(1):18–23, February 2004.

[27] John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, 15(3):285–289, Jun 2005.

[28] John Moult, Krzysztof Fidelis, Burkhard Rost, Tim Hubbard, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)-Round 6. *Proteins*, 61 Suppl 7:3–7, 2005.

[29] I. Muegge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem*, 42(5):791–804, Mar 1999.

[30] J. Novotny, R. Bruccoleri, and M. Karplus. An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol*, 177:787–818, 1984.

[31] Donald Petrey and Barry Honig. Protein structure prediction: inroads to biology. *Mol Cell*, 20(6):811–819, Dec 2005.

[32] B. A. Reva, J. Skolnick, and A. V. Finkelstein. Averaging interaction energies over homologs improves protein fold recognition in gapless threading. *Proteins*, 35(3):353–359, 1999.

[33] John Rowlinson. *Cohesion. A Scientific History of Intermolecular Forces.* Cambridge University Press, 2002.

[34] Michael Rubinstein and Ralph H. Colby. *Polymer Physics.* Oxford University Press, 2003.

[35] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275(5):895–916, 1998.

[36] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1):209–225, 1997.

[37] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–883, Jun 1990.

[38] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7(4):473–501, Aug 1993.

[39] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–362, 1993.

[40] M. J. Sippl. Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, 5:229–235, 1995.

[41] M. J. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *J Mol Biol*, 260:644–648, 1996.

[42] M. J. Sippl and H. Flockner. Threading thrills and threats. *Structure*, 4:15–19, 1996.

[43] M. J. Sippl, M. Hendlich, and P. Lackner. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin beta 4. *Protein Sci.*, 1:625–640, 1992.

[44] M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials from crystal data. 6. determination of empirical potentials for O-H...O=C hydrogen bonds from packing configurations. *J Phys Chem*, 88:6231–6233, 1984.

[45] M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and H. Flockner. Helmholtz free energies of atom pair interactions in proteins. *Fold Des*, 1:289–298, 1996.

[46] M. J. Sippl and S. Weitckus. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, 13(3):258–271, Jul 1992.

[47] Armando D Solis and S. Rackovsky. Improvement of statistical potentials and threading score functions using information maximization. *Proteins*, 62(4):892–908, Mar 2006.

[48] J. M. Thijssen. *Computational Physics.* Cambridge University Press, 1999.

[49] Christian X. Weichenberger and Manfred J. Sippl. NQ-Flipper: Validation and correction of asparagine/glutamine amide rotamers in protein crystal structures. *Bioinformatics (in press)*, 2006.

[50] Christian X. Weichenberger and Manfred J. Sippl. Self-consistent assignment of asparagine and glutamine amide rotamers in protein crystal structures. *Structure (in press)*, 2006.

[51] Markus Wiederstein and Manfred J Sippl. Protein sequence randomization: efficient estimation of protein stability using knowledge-based potentials. *J Mol Biol*, 345(5):1199–1212, Feb 2005.

[52] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*, 285(4):1735–1747, Jan 1999.

[53] Chi Zhang, Song Liu, Hongyi Zhou, and Yaoqi Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci*, 13(2):400–411, Feb 2004.

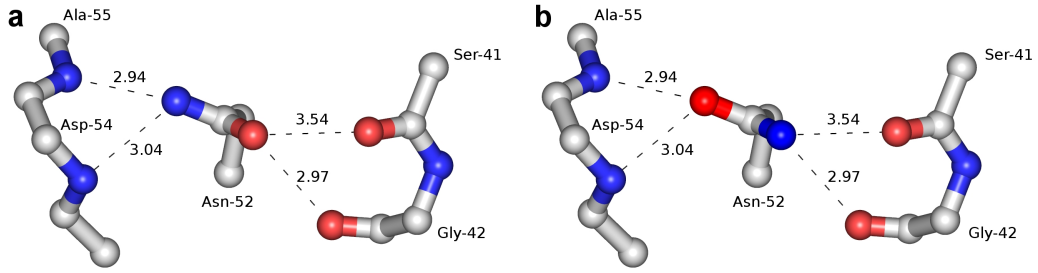

Figure 1: Asparagine Asn-52 of dethiobiotin synthetase (PDB code 1dad; resolution 1.6 Å). Here and in figures 2 and 3 dashed lines and the associated numbers represent distances between atoms (in Å). Atoms are colored by atom type: carbon, grey; oxygen, red; nitrogen, blue. The figures are generated using the program PyMOL (http://pymol.sourceforge.net). **a,** high energy rotamer as found in the crystal structure and **b,** correct low energy rotamer.

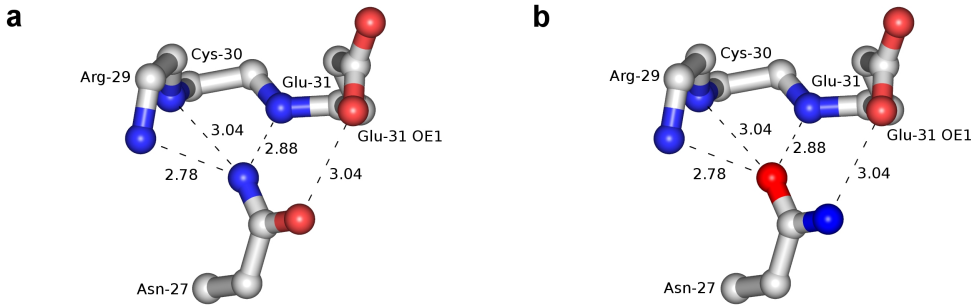

Figure 2: Asparagine Asn-27 of zinc metalloprotease (PDB code 1ezm; resolution 1.5 Å) and the molecular environment of this residues. **a,** high energy rotamer as found in the crystal structure and **b,** correct low energy rotamer.

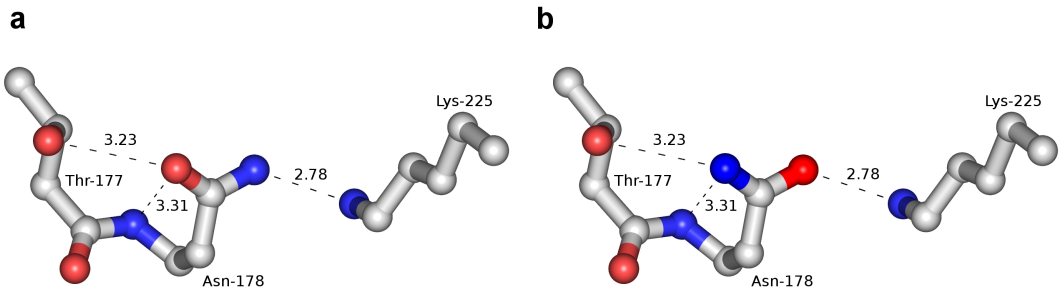

Figure 3: Asparagine Asn-178 of carbonic anhydrase (PDB code 2cba; resolution 1.54 Å). **a,** high energy rotamer as found in the crystal structure and **b,** correct low energy rotamer.

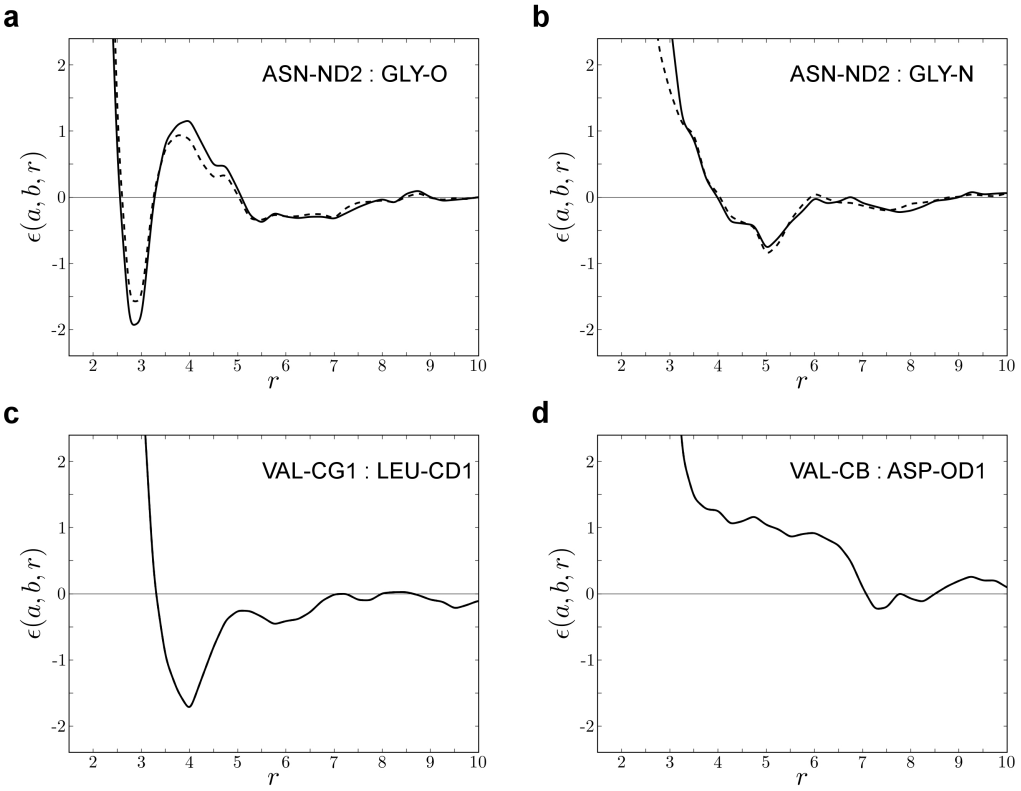

Figure 4: Examples of mean force potentials. Refined (solid lines) and unrefined (dashed lines) potentials of mean force, $\epsilon(a, b, r)$, between atoms of type a and b are plotted as a function of separation r in the distance range $1.5 \leq r \leq 10.0$ Å. All potentials converge to zero at separations larger than $r \approx 15$ Å. The atom types shown are asparagine side-chain amide nitrogen (ASN-ND2), glycine backbone oxygen (GLY-O), glycine backbone nitrogen (GLY-N), valine terminal side-chain carbon (VAL-CG1), leucine terminal side-chain carbon (LEU-CD1), valine β carbon (VAL-CB) and aspartic acid terminal side-chain oxygen (ASP-OD1). **a,** Interaction of the hydrogen bond donor ASN-ND2 and hydrogen bond acceptor GLY-O. **b,** Interaction of the hydrogen bond donor ASN-ND2 and hydrogen bond donor GLY-N. **c,** Interaction of the aliphatic carbon atom VAL-CG1 and the aliphatic carbon atom LEU-CD1. **d,** Interaction of the aliphatic carbon atom VAL-CB and the charged oxygen atom ASP-OD1.

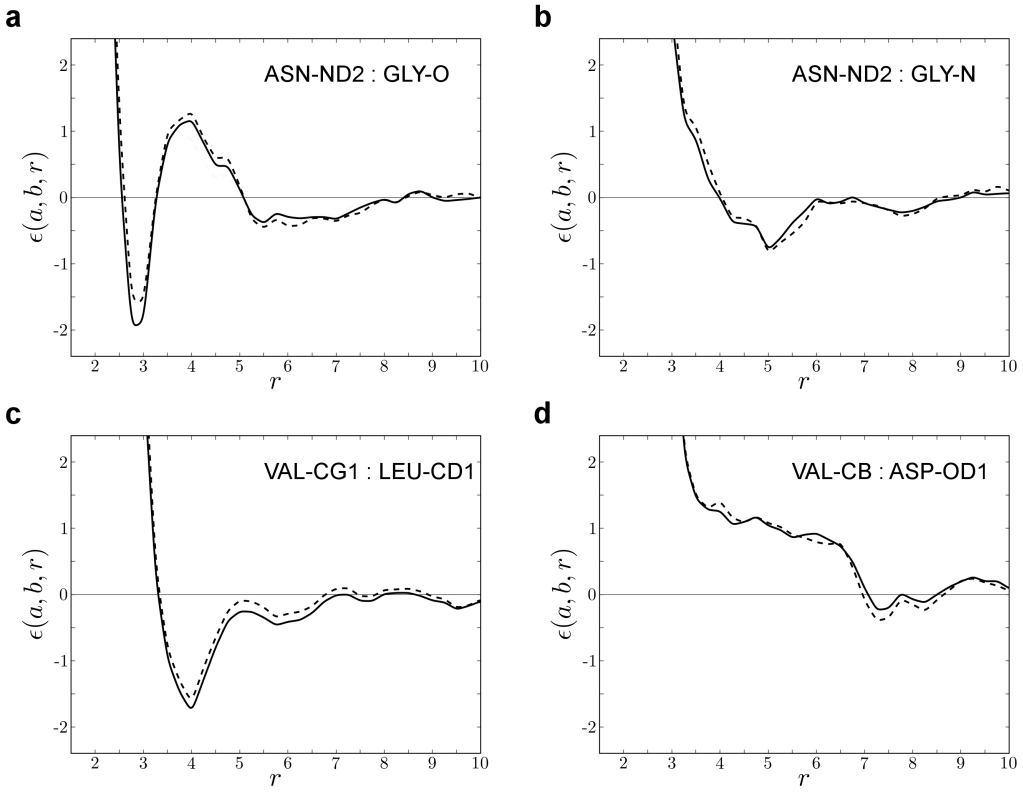

Figure 5: Comparison of mean force potentials compiled from single molecules (dashed lines) and complete crystal structures (solid lines). The atom types shown correspond to those of figure 4.

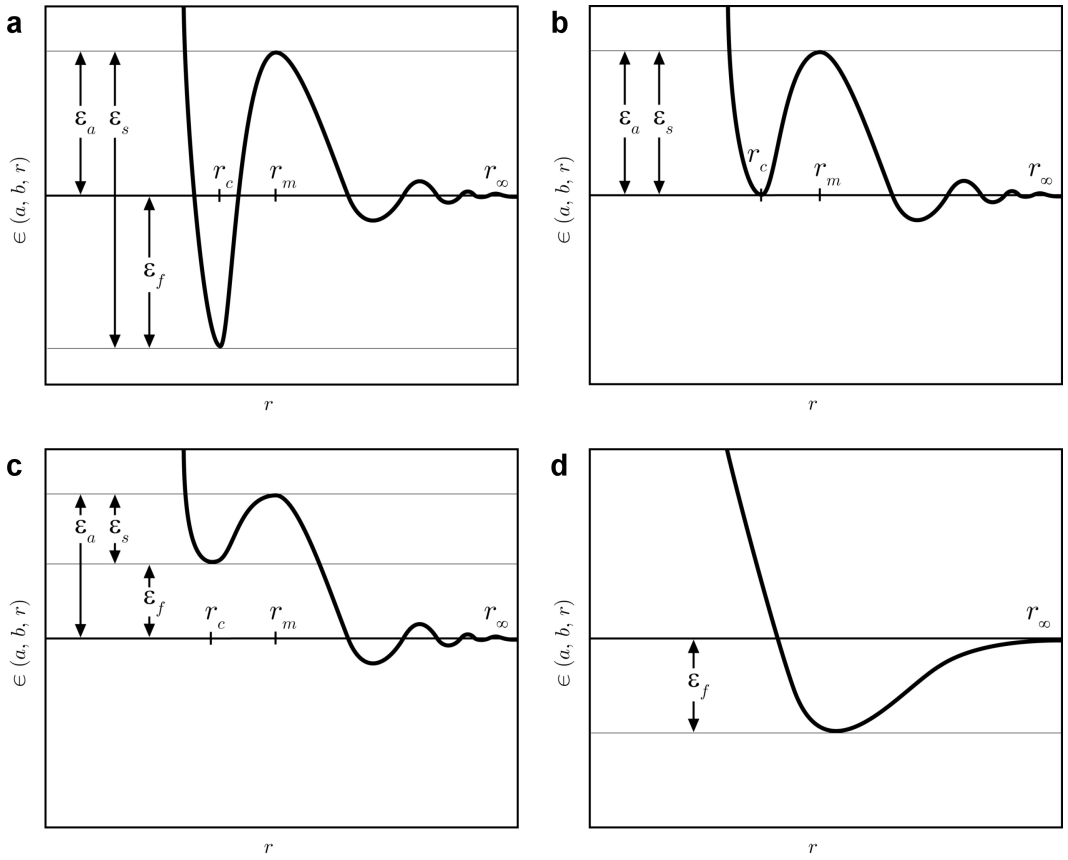

Figure 6: Generic functional forms of mean force potentials found in protein crystal structures. The generic forms of mean force potentials can be characterized by the free energy balance of bond formation, $\epsilon_f = \epsilon(r_c) - \epsilon(r_\infty) = \epsilon(r_c)$, the activation free energy, $\epsilon_a = \epsilon(r_m)$, required to surmount the barrier, and the activation energy of bond disruption, $\epsilon_s = \epsilon_f + \epsilon_a$, required to break the bond. **a,** Molecular lock corresponding to the potential shown in Figure 4 (a). Since $\epsilon_f < 0$, energy is released in the overall process of bond formation. **b,** Special case of a lock where $\epsilon_r = 0$ so that the energy balance of bond formation is zero. **c,** Molecular lock which requires energy input for bond formation (Figure 3 of [50] shows actual examples of this type of interaction). Here bond formation consumes free energy ($\epsilon_f > 0$). **d,** Typical interaction without a barrier corresponding to the interaction of aliphatic atoms of Figure 4 (c). Hydrogen bonds and other highly polar interactions generally have one of the functional forms (a-c), whereas (d) is the typical functional form for hydrophobic interactions.

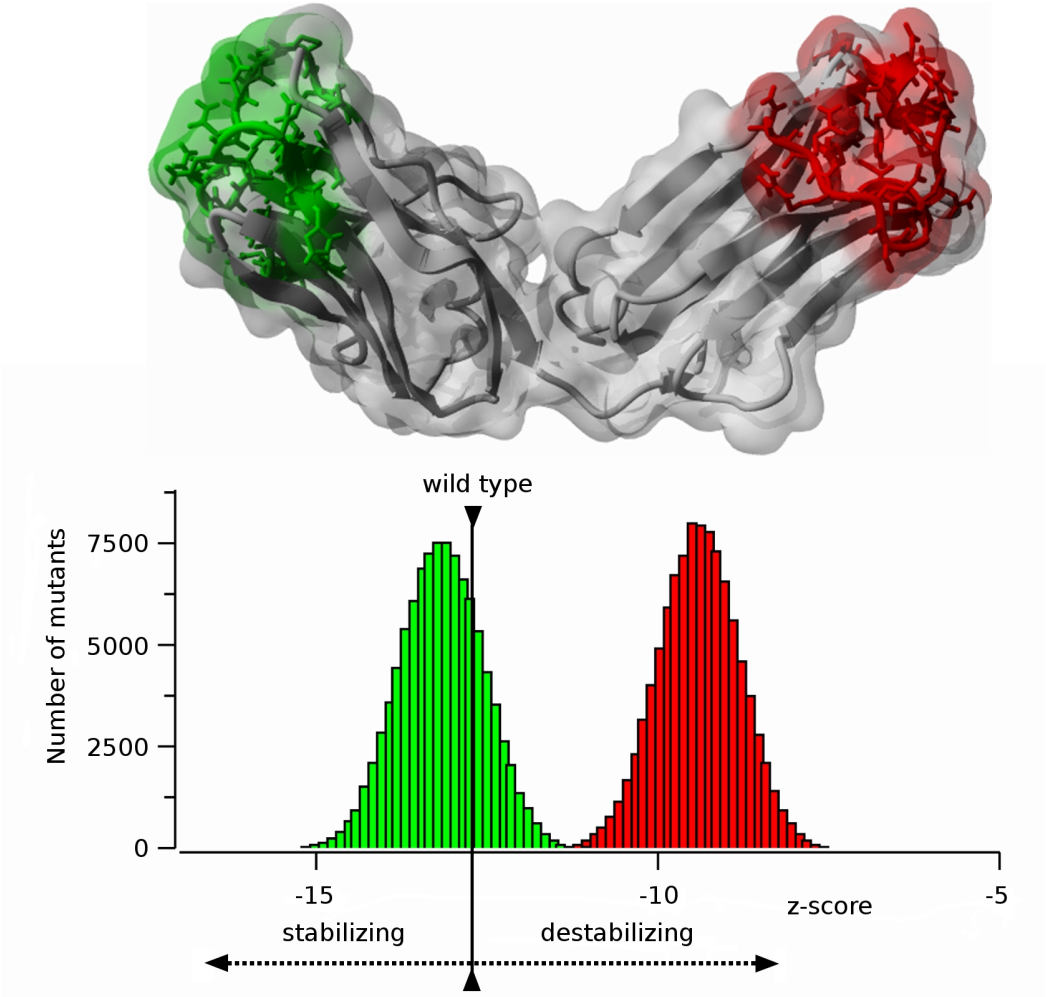

Figure 7: The stability of antibody mutants investigated by potentials of mean force. The complementarity determining regions (CDR) of an antibody are shown in green and an arbitrary region of similar architecture (a 'control' region) is shown in red. The histogram shows the distribution of scores of randomized sequences computed from mean force potentials. More than 70% of the 10^5 randomized CDR sequences stabilize the immunoglobulin fold (green). In contrast, all randomized sequences in the control region destabilize the fold (red) [51].