Statistical approaches to the study of protein folding and energetics

by

Nikolas S. Burkoff

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

MOAC Doctoral Training Centre

July 2014

Contents

Acknowledgements

I would like thank my supervisor, Professor David Wild, for his support and insights throughout the project. I would also like to take this opportunity to thank collaborators I have worked with over the last few years: Dr. Gábor Csányi and Robert Baldock for their work developing the molecular dynamics nested sampling algorithm, Dr. Stephen Wells for his work with rigidity analysis and Dr. John Skilling for general nested sampling and maximum entropy discussions.

I thank the staff of both WSB and MOAC for helping the project run smoothly, particularly Dr. Paul Brown and Sam Mason for their computational support. I would also like to acknowledge the Leverhulme Trust for providing the funding throughout my PhD. Finally, I would like to thank Dr. Csilla Várnai for her hard work and helpful discussions over the last three years and for her calming influence, helping me to continue when the project was going through tough patches. Thank you.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. Unless explicitly stated otherwise, all parts have been composed by myself and have not been submitted in any previous application for any degree. The novel work in this thesis is presented in journal paper format and prefacing each paper there is a summary of my contribution to the work.

Abstract

The determination of protein structure and the exploration of protein folding landscapes are two of the key problems in computational biology. In order to address these challenges, both a protein model that accurately captures the physics of interest and an efficient sampling algorithm are required.

The first part of this thesis documents the continued development of CRANKITE, a coarse-grained protein model, and its energy landscape exploration using nested sampling, a Bayesian sampling algorithm.

We extend CRANKITE and optimize its parameters using a maximum likelihood approach. The efficiency of our procedure, using the contrastive divergence approximation, allows a large training set to be used, producing a model which is transferable to proteins not included in the training set.

We develop an empirical Bayes model for the prediction of protein β-contacts, which are required inputs for CRANKITE. Our approach couples the constraints and prior knowledge associated with β-contacts to a maximum entropy-based statistic which predicts evolutionarily-related contacts.

Nested sampling (NS) is a Bayesian algorithm shown to be efficient at sampling systems which exhibit a first-order phase transition. In this work we parallelize the algorithm and, for the first time, apply it to a biophysical system: small globular proteins modelled using CRANKITE. We generate energy landscape charts, which give a large-scale visualization of the protein folding landscape, and we compare the efficiency of NS to an alternative sampling technique, parallel tempering, when calculating the heat capacity of a short peptide.

In the final part of the thesis we adapt the NS algorithm for use within a molecular dynamics framework and demonstrate the application of the algorithm by calculating the thermodynamics of all-atom models of a small peptide, comparing results to the standard replica exchange approach. This adaptation will allow NS to be used with more realistic force fields in the future.

Chapter 1

Introduction

This introductory chapter is comprised of five sections. The first section introduces the reader to the standard statistical mechanics and thermodynamics concepts used throughout this work. The second section considers the existing techniques for sampling atomistic models. The third and fourth sections discuss protein science and the problem of protein structure prediction, reviewing the current computational protein literature. Finally, the last section summarizes the thesis project, describing the content of the following chapters.

1.1 Statistical Mechanics and Thermodynamics

In this section, standard equilibrium statistical mechanics and statistical thermodynamic concepts used throughout this thesis are introduced and formulae are stated without derivations. The reader is referred to standard texts, for example (1, 2), for a full exposition of the theory.

Thermodynamics is the study of heat and its relationship to energy and work. The theory was developed in the 19th century and the laws of thermodynamics are some of the most elegant and universal throughout science. The theory concerns the macroscopic properties of systems, for example, temperature, pressure and volume, and how they change when heat is transferred into or out of the system.

Equilibrium statistical mechanics seeks to explain equilibrium thermodynamic results as statistical averages of the behaviour of a large number of particles without being concerned with individual particle motions. However, we begin by considering the behaviour of these individual particles. Although a complete description of the behaviour of interacting atoms requires a full quantum mechanical treatment, in many systems, where the electron cloud does not require complete modelling, the potential energy of a system can be well approximated using only the co-ordinates of the atomic nuclei. If such a system is in thermal equilibrium with its surroundings, its behaviour can then be described using the Boltzmann distribution, as discussed below.

1.1.1 Boltzmann Statistics and the Partition Function

A conformation Ω of a set of N atoms (of masses m_i) comprises of a list of N 3-dimensional vectors of atomic co-ordinates. The potential energy, E_Ω, is a function of Ω, capturing the nature of interactions between the atoms. The phase (or conformational) space of the system is defined as the set of all possible Ω and the potential energy surface (PES) is defined as the $3N$-dimensional function E_Ω.

Each atom is endowed with a momentum vector $\mathbf{P} = \{\mathbf{p_1}, \ldots, \mathbf{p_N}\}$ and the total energy of the state $\{\Omega, \mathbf{P}\}$ is defined as the sum of its potential and kinetic energies:

$$E(\Omega, \mathbf{P}) = E_\Omega + \sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2m_i}.$$

If the system is restricted to a volume V and in thermal equilibrium with its surroundings at temperature T, then it behaves according to Boltzmann statistics at inverse temperature $\beta = 1/k_B T$ where k_B is the Boltzmann constant ($\approx$ 2x10^{-3} kcal/mol/K). The probability of finding the system in state $\{\Omega, \mathbf{P}\}$ is then given by

$$\begin{aligned} \mathbb{P}(\{\Omega, \mathbf{P}\}) &\propto \exp(-E(\Omega, \mathbf{P})\beta) \\ &= \exp\left(-\beta \sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2m_i}\right) \exp(-E_\Omega \beta) \end{aligned}$$

and the probability of the system being in conformation Ω is thus given by

$$\mathbb{P}(\Omega) \propto \exp(-E_\Omega \beta).$$

The (configuration) partition function $Z(\beta)$ is defined as the normalization constant of this conformational distribution,

$$Z(\beta) = \int \exp(-E_\Omega \beta) d\Omega,$$

where the integral is over the 3N-dimensional conformation space. Following standard statistical theory, the expectation of any function $A(\Omega)$ is given by

$$\langle A \rangle_\beta = \frac{1}{Z(\beta)} \int A(\Omega) \exp(-E_\Omega \beta) d\Omega.$$

By setting $A = E_\Omega$, the above formula yields the expectation of potential energy, named the internal (or thermodynamic) energy, which is often denoted by $U(\beta)$.

Connection to Bayesian Statistics The Boltzmann distribution can be recast in the Bayesian probability framework by defining a likelihood function $L(\Omega) = \exp(-E_\Omega)$, and using a prior, $\pi(\Omega)$, uniform over the conformational space. The posterior distribution $P(\Omega) \propto L(\Omega)\pi(\Omega)$ then corresponds to the Boltzmann distribution at inverse temperature $\beta = 1$, and the normalization constant,

$$Z = \int L(\Omega)\pi(\Omega) d\Omega,$$

also known as the marginal likelihood or 'evidence' in Bayesian terminology, is precisely the thermodynamic partition function at the same temperature. Replacing the likelihood with $L^\beta(\Omega)$ gives the Boltzmann distribution at inverse temperature β. This correspondence allows algorithms and techniques developed for Bayesian inference to be used with atomistic systems behaving according to Boltzmann statistics.

1.1.2 Statistical Thermodynamics

In the previous section we looked at how the *microscopic* properties of a system change with changing temperature, namely the probability of the system taking a specific conformation. In this section we are concerned with the *macroscopic* properties of a system, such as pressure and heat capacity. Unlike microscopic properties, macroscopic properties can often be measured experimentally, so when comparing computer simulations to experiments, their calculation is crucial. As shown below, the partition function, $Z(\beta)$, is of fundamental importance in linking the microscopic and macroscopic scales. For example, the average energy U is given by $-\partial \log Z/\partial\beta$.

A macroscopic state is a set of microscopic conformations. For example, in the case of water we could define the macroscopic states of 'ice', 'liquid water' and 'water vapour'. 'Ice' would consist of all conformations Ω, for which the water molecules are placed in a regular crystal lattice. A natural question is, given that the system has temperature T, what is the probability the system is found in a particular macrostate? In the case of a system with two macrostates, X and Y, using the Boltzmann distribution we find

$$\mathbb{P}(X) = \frac{\int_{\Omega\in X} \exp(-E_\Omega\beta)d\Omega}{Z(\beta)}$$

and the (Helmholtz) free energy of state X, F_X, is given by

$$F_X = -\frac{\log\left(\int_{\Omega\in X} \exp(-E_\Omega\beta)d\Omega\right)}{\beta}.$$

The probability of state X is thus given by

$$\frac{1}{1+\exp(-\beta(F_Y - F_X))},$$

where F_Y is the free energy of state Y.

The free energy difference, $F_Y - F_X$, determines the relative probabilities, and due to the exponential in the formula, if the difference is large then the system will be found in the state with smallest free energy with a probability of effectively 1. Hence the system minimizes its free energy.

Macroscopic states do not have to have categorical labels. For example, macrostates could be labelled by the real number d, where d is the distance between two specific atoms. The chosen labels are known as reaction co-ordinates. The calculation and visualization of the free energy landscape – that is, the function which maps reaction co-ordinates to their free energies at a given temperature – can be an important step in understanding the thermodynamics of the system (3).

The absolute free energy of the whole system, $A = -\log Z/\beta$, is an important thermodynamic variable and it can be shown that $A = U - ST$, where S is the entropy of the system, given by $-\partial A/\partial T$ when the volume of the system and its number of particles are kept fixed.[1]

Intuitively, a macroscopic state which comprises of a larger number of conformations has a higher entropy than a state which is comprised of fewer conformations. Returning to our two-state system, let Y correspond to liquid water and X to ice. In this case, Y is comprised of a large number of conformations of (relatively) high energy and thus has high entropy and high internal energy. Conversely X is comprised of a (relatively) small number of conformations of low energy and thus has low entropy and low internal energy.

[1] As with all thermodynamic variables, many equivalent definitions exist. The one given here relates to the constant volume and number of particles formulation that has been followed thus far and hence does not require the introduction of further terms.

In thermal equilibrium at low temperatures, $F_i \approx U_i$ and hence $F_X \ll F_Y$ and we have ice. However, at high temperatures, as $F_i = U_i - TS_i$, entropic effects are more important and hence $F_Y \ll F_X$ and the system is liquid. The special temperature of 0°C corresponds to the case $F_X = F_Y$ and the system is said to undergo a phase transition at this temperature. Phase transitions are discussed further in the next section.

The (constant volume) heat capacity

$$C_v = \frac{\partial U}{\partial T} \equiv k_B \beta^2 \frac{\partial^2 \log Z}{\partial \beta^2} \equiv k_B \beta^2 (\langle E_\Omega^2 \rangle_\beta - \langle E_\Omega \rangle_\beta^2)$$

corresponds to the amount of energy that must be input into the (fixed volume) system in order to increase its temperature. Two equivalent definitions are shown, one a derivative of the logarithm of the partition function, the other closely related to the fluctuations of the potential energy of the Boltzmann distributed system. The heat capacity is of particular interest as it can be directly measured experimentally.

Other thermodynamic variables of interest can also be calculated, given the free energy (or equivalently the logarithm of the partition function), for example the pressure of the system is given by $p = -\partial A/\partial V$ when the temperature and number of particles are kept fixed.

Other Thermodynamic Distributions

The Boltzmann distribution and thermodynamic formulae described above assume the number of particles (N), the total volume (V) and the temperature (T) of the system are all kept constant. This distribution is often called the canonical or NVT distribution. By changing the quantities kept constant, other important thermodynamic distributions can be derived and examples are shown in Table 1.1. Although it is possible to perform experiments keeping N, V and T fixed, for example by using a bomb calorimeter, it is much more typical to keep the pressure rather than volume constant, hence the importance of the isothermal-isobaric ensemble.

For a system in isothermal-isobaric equilibrium, analogous thermodynamic variables to those described above can be defined. For example, the system minimizes its Gibbs free energy $G = U + pV - TS$. Analogous to the internal energy U is the enthalpy $H = U + pV$, and the (constant pressure) heat capacity, C_p, is defined as $\partial H/\partial T$. For the systems we consider, where volume per molecule is small and we are at low (atmospheric) pressure, pV is negligible compared to U and so $H \approx U$ and $G \approx A$.

Distribution	Constants	Partition Function	State Function
Microcanonical	N,V,E	W (# states)	Entropy $S = k_B \log W$
Canonical	N,V,T	$Z = \int \exp(-E_\Omega \beta) d\Omega$	Helmholtz free energy $A = -\log Z/\beta = U - ST$
Grand canonical	μ,V,T	$\mathcal{Z} = \sum_{N_i} Z \exp(N_i \mu \beta)$	Grand potential $-\log \mathcal{Z}/\beta$
Isothermal-isobaric	N,p,T	$\Delta \propto \int Z \exp(-\beta PV) dV$	Gibbs free energy $G = -\log \Delta/\beta = A + pV$

Table 1.1: A list of the common thermodynamic distributions, their constants, partition functions and macroscopic state functions. μ is the chemical potential defined as $\partial A/\partial N$ when volume and temperature are kept fixed.

1.1.3 Phase Transitions

From an intuitive perspective, phase transitions are abrupt changes of macroscopic properties of a system with changing external conditions. For example, the decrease in temperature of a liquid can lead the system to freeze into a solid, and lowering the temperature of a paramagnetic material below its Curie point causes an abrupt transformation from paramagnetism to ferromagnetism, i.e. having a magnetic moment. Phase transitions also occur in more abstract branches of science; for example, in random graph theory, as the probability of edges increases above a critical threshold, the probability of there being a single connected component abruptly approaches 1. Mathematical theories, for example Mean–field theory and Ginzburg-Landau theory, have been developed in order to explain the behaviour of systems at phase transitions and we refer the reader to standard texts for full descriptions (2, 4).

From a (constant NVT) statistical mechanics perspective, a phase transition occurs when the Helmholtz free energy is non-analytic. If the first derivative of A is discontinuous, then the phase transition is first order, whereas if the first derivative is continuous but its second derivative is discontinuous, then the transition is described as second order. Examples of first order transitions include the melting of ice and boiling of water. Examples of second order phase transitions include the ferromagnetic transition from paramagnetism to ferromagnetism. Figure 1.1 shows the behaviour of the internal energy and heat capacity of systems with first and second order phase transitions. Analogous behaviour can be found for systems in the isobaric-isothermal distribution with G replacing A.

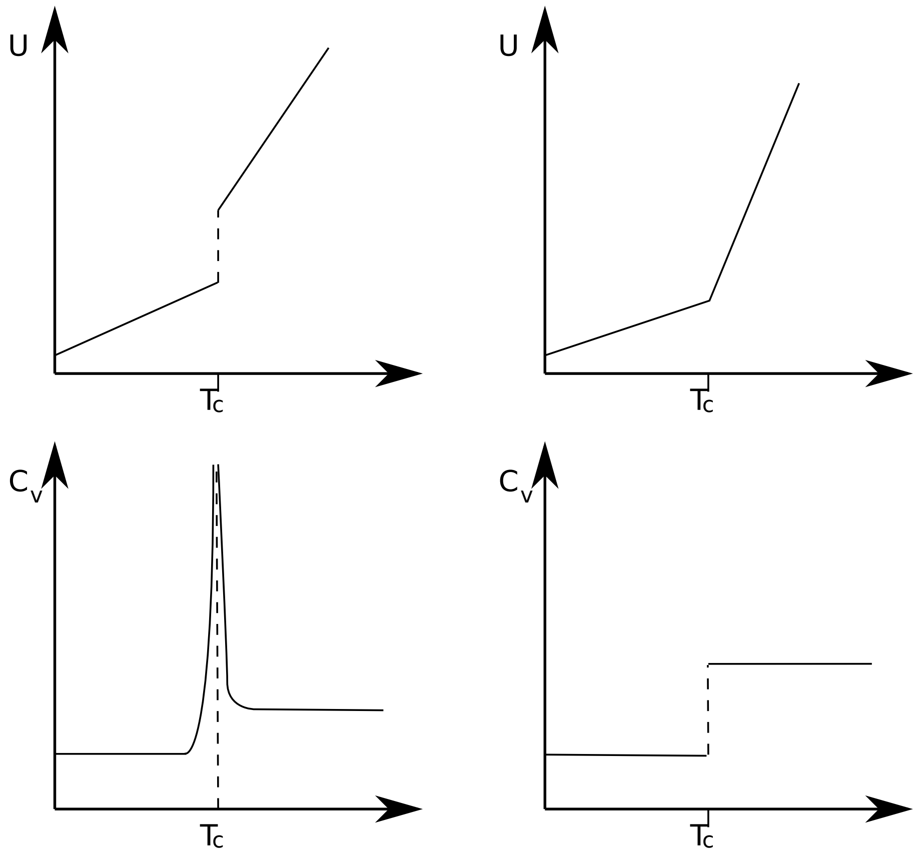


Figure 1.1: Left: Behaviour of the internal energy (top) and heat capacity (bottom) of a system with a first order phase transition at temperature T_c. Right: Behaviour of the internal energy (top) and heat capacity (bottom) of a system with a second order phase transition at temperature T_c.

The phase transitions described above involve bulk systems with enormous numbers of degrees of freedom, yet these systems cannot be modelled computationally in atomic detail. Only the behaviour of smaller (or periodic) systems with fewer degrees of freedom can be studied computationally. These systems, or indeed any finite system, cannot exhibit a true first order phase transition, with a genuine discontinuity in internal energy. They can, however, exhibit a quasi-first order transition, where the C_v curve is sharply peaked but remains finite and the U versus T graph is sigmoidal rather than discontinuous (Figure 1.2 (left)). Typically as $N, V \to \infty$ with N/V remaining constant, the transition becomes 'sharper', i.e. the C_v peak is higher and occurs over a shorter temperature range, and rapidly approaches a true first order transition in the thermodynamic limit.

Finally, we consider how the probability distribution of the potential energy behaves at a first order phase transition (5). Firstly, away from the phase transition, it can be shown that the distribution is Gaussian, with mean $U(T)$ as shown by Figure 1.2 (a,e). Near the critical temperature, T_c, the distribution is no longer unimodal and non-negligible probability mass is found around two separate energy values U_X and U_Y, the average energy of the system at this temperature, given that the system is restricted to phase X or respectively Y. This is shown by Figure 1.2 (b,d). At T_c the probability of being found in phase X is exactly a half as shown in Figure 1.2 (c). At T_c the free energy of the small number of low energy states of phase X exactly balances the free energy of the large number of high energy states of phase Y. As we show below, this behaviour causes serious difficulties when computationally sampling from such a system.

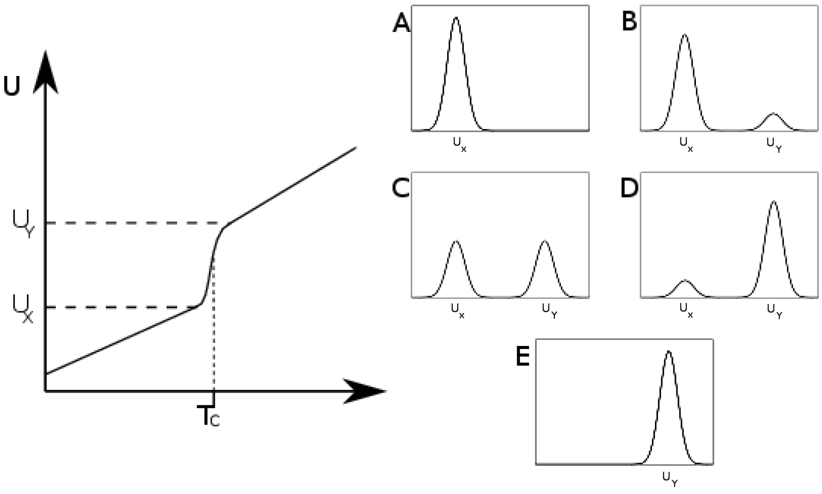


Figure 1.2: Left: The internal energy (U) as a function of temperature for a system exhibiting a quasi–first order transition at temperature T_c. Right: The potential energy probability distribution for this system at five different temperatures A) $T \ll T_c$ B) $T < T_c$ C) $T = T_c$ D) $T > T_c$ E) $T \gg T_c$

1.2 Sampling of Atomistic Systems

Given a molecular model and potential energy function E_Ω, we are often interested in calculating kinetic and thermodynamic properties of the system, in order to compare to an experiment, to compare to other, often similar, systems or to provide new insight into how the system behaves when changing its external conditions. Kinetic, or dynamic, properties of the system, such as diffusion and equilibration rates, are time-dependent properties, whereas thermodynamic averages, such as free energies, are time-independent properties of the system when it is in equilibrium. In order to calculate these properties, it is necessary to evolve the state of the system and draw samples for analysis.

In this section we introduce the important sampling techniques of molecular dynamics and Monte Carlo sampling. We then describe the *sampling problem* associated with sampling from Boltzmann distributions and discuss methods which attempt to solve this problem. We end the section with a description of nested sampling, a new sampling technique developed for Bayesian inference, which has shown potential for improving the sampling of atomistic systems.

However, we first briefly mention molecular models and their energy functions. A model of a molecule consists of a series of atoms and information as to how they are covalently bonded, together with a set of parameters such as atomic charges, bond lengths and valence angles. A model of a system consists of a set of molecule models together with details of boundary conditions, system volume, size and shape.

The potential energy function of physical models comprises of three separate terms

$$E_\Omega = E_\Omega^{\text{bonded}} + E_\Omega^{\text{non-bonded}} + E_\Omega^{\text{external}}.$$

E_Ω^{bonded} captures the energetic interactions between covalently bonded atoms. These interactions include bond stretching, usually modelled with a harmonic function, valence angle bending and bond rotations. $E_\Omega^{\text{non-bonded}}$ captures interactions between non-covalently bonded atoms, for example, electrostatic interactions, using a Coloumb potential, and van der Waals interactions, often modelled by a Lennard Jones (LJ) potential.[2] $E_\Omega^{\text{external}}$ consists of terms external to the set of atoms, such as the energy from an external magnetic field. We discuss specific models for proteins in §1.3.4 and the reader is referred to (6) for more details of general molecular models.

1.2.1 Molecular Dynamics Sampling

First developed computationally in the late 1950s by Alder and Wainwright (7), molecular dynamics (MD) evolves the state of the system by integrating Newton's equations of motion. Given the potential energy function E_Ω, initial atomic co-ordinates $\Omega(0) = \{\mathbf{x}_i(0)\}$ and velocities $\{\mathbf{v}_i(0)\}$, we have

$$m_i \frac{d^2\mathbf{x}_i}{dt^2} = \mathbf{F_i},$$

where the forces $\mathbf{F_i}$ are given by $-\nabla_{\mathbf{x}_i} E_\Omega(\mathbf{x})$. Solving this system of differential equations leads to a trajectory $\Omega(t)$.

It can be shown that the total energy (E_Ω+kinetic energy) of the trajectory remains constant, and therefore, assuming the ergodic hypothesis described below is satisfied, the samples generated are distributed according to the microcanonical distribution.[3]

When integrating the equations numerically, care must be taken to use an integration scheme which does not introduce large discretization errors. Schemes such as the Verlocity Verlet, described in (8), are widely used.

In the thermodynamic limit of a large number of particles, the canonical and microcanonical ensembles coincide. However, for small systems, in order to sample their canonical distributions, a thermostat must

[2] If atomic distance is r, then the LJ potential is given by $4\epsilon((\sigma/r)^{12} - (\sigma/r)^6)$ where ϵ and σ are parameters. The LJ potential models both short-range repulsion and the long-range attraction due to the fluctuating charge densities of induced dipoles.

[3] Technical details have been omitted. For example, in non-periodic systems the angular momentum, J, is also a constant of motion and samples come from a 'NVEJ' distribution.

be used to ensure constant temperature. The Andersen thermostat (9) is a simple procedure, coupling the system to an external heat bath to maintain the desired temperature. After each timestep, each atom undergoes a collision with the heat bath with probability ν. A collision involves resampling the momentum of an atom, with its new momentum sampled from the Maxwell-Boltzmann distribution[4] at the desired temperature. It can be shown that the samples generated using this procedure come from the canonical distribution.

Other schemes, such as Langevin dynamics or the Nosé–Hoover thermostat, have also been developed. For further details on these and sampling atomistic systems in general, we refer the reader to the excellent discussion in (10). Barostats, which control the pressure of the system by changing its volume, have also been developed, enabling isothermal-isobaric trajectories to be computed.

When using MD trajectories to estimate kinetic properties of the system, for example diffusion rates, it is important to choose a thermostat and barostat which preserve the kinetics of the system. The resampling of momenta in the Andersen thermostat described above means it cannot be used to estimate time-dependent system properties.

As well as the calculation of kinetic properties, samples from MD trajectories can be used to estimate thermodynamic averages. For example, given an NVT trajectory $\Omega(t), \ldots, \Omega(Mt)$, the thermodynamic energy can be estimated as

$$\langle E_\Omega \rangle_\beta \approx \frac{1}{M} \sum_{i=1}^{M} E_{\Omega(it)}$$

and this will, in the limit of large M, converge to the ensemble average, assuming the ergodic hypothesis. The ergodic hypothesis, originally formulated by Boltzmann at the end of the 19th century, proposes that the long-time average of a property over a trajectory does indeed converge to its ensemble average. Whilst usually assumed true, there are cases when, in practice, it does not hold, such as when the system has long-lived metastable states.

1.2.2 Monte Carlo Sampling

An alternative to MD simulation is that of Markov Chain Monte Carlo (MC) sampling. The MC method was devised in the 1940s and '50s for use on some of the very first computers. We refer the reader to (11) for an interesting historical perspective on the invention of MC methods.

The basic idea of this method is to devise a Markov chain,[5] such that its equilibrium distribution is the distribution from which we wish to sample. Again, assuming ergodicity, ensemble averages can then be estimated from the MC samples.

We describe the Metropolis-Hastings algorithm (12), a generalization of the original Metropolis procedure (13). In order to use this method, a set of moves which evolve the state of the system must be devised. We define $\mathbb{P}(\Omega \to \Omega')$ as the probability that, given conformation Ω, a move is chosen which would transform it into conformation Ω'. Assuming the system is in conformation Ω and we have proposed to evolve the chain to conformation Ω', we accept this move with the Metropolis-Hastings acceptance probability

$$\min\left(1, \frac{P(\Omega')\mathbb{P}(\Omega' \to \Omega)}{P(\Omega)\mathbb{P}(\Omega \to \Omega')}\right),$$

where $P(\Omega)$ is the distribution of interest. In the case of the canonical distribution, $P(\Omega')/P(\Omega) =$

[4] The marginal distribution of momenta when the system follows the canonical distribution.

[5] A Markov Chain is a sequence of random variables $X_1, X_2, \ldots$ such that $\mathbb{P}(X_k | X_{k-1}, X_{k-2} \ldots, X_1) = \mathbb{P}(X_k | X_{k-1})$. In atomistic systems X_i is a conformation Ω_i.

$\exp(-\beta(E'_\Omega - E_\Omega))$. If we reject the move, the MC chain remains in state Ω and another move is proposed. The equilibrium distribution of this chain is $P(\Omega)$.

The choice of the move set is an important factor in MC procedures. Since only $E'_\Omega - E_\Omega$ is required in order to accept or reject moves, local moves, that is those which keep most of the system fixed, are usually more efficient; only the part of the potential energy function which involves the atoms that have moved needs to be calculated for each iteration. Also, at low temperatures, large changes to the system will likely be rejected. In atomistic systems, moves typically include molecular translations and rotations and intra-molecular bending and rotation of covalent bonds.

Alongside canonical sampling, the MC procedure has been used to sample other thermodynamic distributions, as described in (10). These even include the grand canonical ensemble, where proposal moves include the addition or removal of atoms from the system.

Unlike MD sampling, samples generated from an MC sampler cannot be used to estimate kinetic properties of the system but, if appropriate moves exist, MC sampling can be a very efficient way of exploring conformation space.

1.2.3 Advanced Sampling Techniques

Unfortunately, in many cases standard MC or MD sampling is not effective. For example, in a multimodal system with low probability of swapping between modes at temperature β, the system will become stuck in a single mode and require an enormous runtime in order to produce accurate ensemble averages.

If we are using MD or MC to sample a system with a phase transition near the critical temperature, for example Figure 1.2 (right C), it will be nearly impossible to equilibrate the samples between the (relatively) small number of low energy samples with energy around U_X and the large number of high energy samples with energy near U_Y. This problem can occur in both multi and unimodal systems. We return to this example in more detail below.

Furthermore, if the partition function is required, it is possible to use the samples, $\Omega_1, \ldots \Omega_M$ to produce an estimate

$$Z^{-1} = \langle \exp(E_\Omega \beta) \rangle_\beta \approx \sum_{i=1}^{m} \exp(E_{\Omega_i} \beta).$$

However, this estimate, the 'harmonic mean approximation', has infinite variance at low temperatures (14) and hence should not be used.

In order to avoid these problems, more sophisticated sampling algorithms involving MC, MD or a combination of both techniques have been designed. In this section we discuss some of these algorithms before introducing a novel algorithm, nested sampling, which has the potential to significantly improve the sampling of atomistic systems.

General Sampling Algorithms

Although technically an optimization algorithm, we first describe simulated annealing (SA), as the ideas it introduces are applicable to the sampling algorithms described below. SA was developed by Kirkpatrick and colleagues in the early 1980s (15). The algorithm starts by exploring the conformation space using a high temperature canonical MC chain. Gradually throughout the simulation, the temperature is lowered, following a chosen schedule, until we are sampling from the canonical distribution of the temperature of interest.

At high temperatures (low β), the factor $\exp(-E_\Omega \beta)$ is less important: in the limit of high temperature this term is constant with respect to E_Ω, and exploring is easy, as it is uniform in conformational space. At the start of SA, the chain can explore the conformation space more easily and thus larger MC steps can be taken. As the temperature is decreased this exponential factor becomes more important and smaller MC steps must be taken. The system is slowly cooled, hopefully resulting in finding the correct (free) energy minimum of the system. The slow cooling gives the a system chance to properly equilibrate.

Once the temperature has been reduced, the conformation will eventually become stuck in a local minimum. An adaptation of SA, the simulated tempering algorithm (16), allows increases of the temperature of the system, so that the conformation is no longer trapped. Thus, when the temperature is lowered again, the system can explore another minimum.

One of the most widely used sampling algorithms is that of parallel tempering (PT). It was first developed by Swenson in 1986 (17) and, like SA, the temperature of the system becomes a parameter of the algorithm. In the PT procedure, a set of separate MC chains (replicas) are run at different temperatures in parallel. Periodically one proposes that the temperatures of two chains are swapped. This proposal is accepted using an acceptance criterion similar to the Metropolis-Hastings criterion described above.

A key advantage of this approach is that multimodal systems can be sampled much more efficiently. Low temperature replicas, which would have been be trapped in local modes, can 'escape' by swapping with a conformation from a chain with higher temperature. The high temperature replicas can themselves more easily escape modes, as explained above.

The algorithm outputs samples from the canonical distribution at a variety of temperatures which can be used to produce thermodynamic averages over a range of temperatures. Thermodynamic estimates for other temperatures can also be estimated from these samples using a procedure such as Boltzmann reweighting (18).

However, care must be taken to ensure enough replicas are used (order $\sqrt{N}$ for system size N), so that the probability of accepting proposed temperature swaps is not too small. Also, the set of temperatures used must be chosen with care. For example, at a phase transition, more replicas are needed near the critical temperature to ensure proper equilibration, even though the existence and location of the transition is not known *a priori*.

Replica exchange molecular dynamics (REMD) (19) is the same algorithm as PT, except that instead of each replica running an MC chain, it follows an MD trajectory. Care must be taken to rescale the velocities after replicas have been exchanged. REMD affords the benefits of PT for systems for which MC moves are not as efficient as MD.

REMD and PT are widely used algorithms, and various adaptations and improvements have been developed, for example, adapting the temperature of replicas throughout the simulation in order to maximize efficiency (20). For a general overview of PT and REMD, consult (21).

Alternative ways of combining MC and MD sampling have also been researched. As an example, we describe the Hybrid Monte Carlo (HMC) algorithm of Duane *et al.* (22). The HMC algorithm is a special MC chain, whereby proposal conformations are generated by giving the atoms of the current conformation momenta and then running a short MD trajectory.

Perhaps counterintuitively, rather than sampling the canonical distribution directly, it may be more efficient to sample from a non-canonical distribution and then weight samples to estimate canonical thermodynamic averages. Algorithms which follow this approach are named biased or extended ensemble sampling methods. As an example we describe the multicanonical sampling algorithm (23). For this method, instead of attempting to sample conformations $\propto \exp(-E_\Omega\beta)$, we sample $\propto 1/g(E_\Omega)$ where $g(E) = \int_\Omega \delta(E_\Omega - E)d\Omega$ is the density of states (δ is the Dirac delta function).

This choice of distribution implies that the histogram of potential energies of the samples will be flat, so that the system is sampling 'uniformly in energy'. This has the advantage of removing the exponential barrier between different modes which is found with canonical sampling algorithms. However, before being used, the algorithm requires a learning stage in order to estimate $g(E_\Omega)$ and, as the system size increases, the number of parameters which need to be estimated in order to resolve $g(E_\Omega)$ accurately can cause problems.

The challenge of estimating $g(E_\Omega)$ can be overcome by a powerful non-Markovian (history-dependent) algorithm, the Wang-Landau sampling algorithm (24), which interactively changes the Markov acceptance criteria to ensure that the system converges to the $1/g(E_\Omega)$ distribution.

As the exploration of phase space is such an important problem both in atomistic systems and statistics in general, many other sampling algorithms and combinations of existing algorithms have been developed. Examples include equi-energy sampling (25) and the 1/k ensemble algorithm (26), a variant of the multicanonical sampling algorithm, notable in this work because it aims to sample conformations uniformly in log phase-space volume rather than energy. We now, however, turn our attention to sampling algorithms specifically developed for calculating the partition function or free energies of atomistic systems.

Partition Function and Free Energy Estimation Algorithms

Within the statistical literature, algorithms have been developed for calculating the marginal likelihood (partition function) of models. Examples include annealed importance sampling (27), Laplace's method (28) and bridge and path sampling (29). Interesting discussions from a statistical viewpoint can be found in (29, 30). Here, however, we focus on methods developed for atomistic systems. Rather than specific values of free energies themselves, statistical physicists are typically more interested in free energy differences, either between two separate systems or between macrostates within the same system.

Unlike thermodynamic variables such as the heat capacity, which can be calculated directly as ensemble averages, free energies are related to volumes in conformation space (consider the definition given in §1.1.2). This makes their calculation more challenging.

One of the most common methods for calculating free energy differences is that of thermodynamic integration (TI) (31). As a simple example of TI, consider the case of two systems X and Y, identical apart from the fact that X has potential energy function E_Ω^X and Y has potential energy function E_Ω^Y. Introducing a parameter λ and defining a new potential energy function

$$E_\Omega^\lambda = (1-\lambda)E_\Omega^X + \lambda E_\Omega^Y$$

it can be shown that the free energy difference between systems X and Y is given by

$$\int_0^1 d\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda .$$

The integrand is an ensemble average which can be estimated through MC or MD sampling for different values of λ, and numerical integration can then be used to estimate the free energy difference.

Often, it is of interest to calculate how the free energy of a system changes when a reaction co-ordinate, a chosen function of the atomic co-ordinates such as an atomic distance or angle, is used. The free energy surface along such a co-ordinate is known as the potential of mean force (PMF) (32). PMFs can be challenging to compute if the surface contains high free energy barriers.

Umbrella sampling (33) is traditionally used to calculate PMFs. In this method, an extra biasing potential is used to force the system to sample a chosen value of the reaction co-ordinate. The procedure is then repeated for different values of the reaction co-ordinate and reweighting these samples generates a more accurate PMF compared to naive MD or MC sampling. In order to further improve the efficiency of the procedure, more sophisticated reweighting techniques, such as the weighted histogram analysis method (WHAM), have been applied to the calculation of PMFs (34).

There have also been many other approaches to the calculation of free energies, including thermodynamic perturbation (35), approximate rapid methods such as λ-dynamics (36) and more specialist algorithms specifically for the calculation of free energies of solids (37).

Nested Sampling

Nested sampling is a novel sampling algorithm invented by Skilling in 2004 in order to estimate the marginal likelihood (i.e. the partition function) in Bayesian inference problems (38, 39). The algorithm is designed for models for which the bulk of the posterior probability mass is contained in an exponentially small area of the prior space. It is particularly efficient at sampling systems which undergo first order phase transitions. The algorithm has been used in a wide variety of fields, including astrophysics (40), system biology (41) and bioinformatics (42).

As shown in §1.1.1, Bayesian inference algorithms can be applied to atomistic systems, and nested sampling has been shown to be an order of magnitude more efficient at calculating heat capacity curves of small Lennard Jones clusters compared to parallel tempering (43). The nested sampling algorithm is *athermal*, and when run, produces samples from throughout the *whole* potential energy surface of the system. These samples can be used to estimate thermodynamic variables at any temperature. Pártay *et al.* have used these samples to generate energy landscape charts, which give a high-level visualization of the potential energy surface. They have also used these charts to define macrostates of the system without having to pick specific reaction co-ordinates (43).

A simple description of the algorithm follows below and more details and in-depth discussions can be found in Chapters 2, 5 and 6 of this work. The algorithm begins by choosing K samples uniformly distributed according to the prior distribution, which in the case of atomic systems is uniform over conformation space. These samples are the current 'active set' and the energy of each sample is calculated.

The conformation with the highest energy is removed from the set and its energy, E_1 is saved. A new conformation, uniformly distributed over the set of conformations with energy less than E_1 is generated and the procedure continues, generating a series of energy levels $E_1, E_2, E_3 \ldots$.

The algorithm does not prescribe how exactly to generate a sample uniformly distributed with energy below E_i. One method involves making a copy of one of the conformations in the active set and then running a MC chain to move the copy away from its starting point. In order to sample uniformly, the standard Metropolis-Hastings acceptance criterion is replaced by a simple criterion: accept the move *iff* the energy remains below the current energy cutoff, assuming the moves are such that the probability of proposing $\Omega \rightarrow \Omega'$ and $\Omega' \rightarrow \Omega$ are the same.

As explained below, one of the key advantages of nested sampling is that the phase space is shrunk by a constant factor, α, at every iteration. By using the expected value of the shrinkage ratio, we can estimate $\alpha \approx \exp(-1/K)$. Estimates for α which take into account the statistical uncertainties of the procedure can also be used.

By defining $X_i = \alpha^i$, energy level E_i and the conformation removed at this energy level (Ω_i) represent a fraction $\omega_i = X_{i-1} - X_i$ of the prior (conformational) space and, at inverse temperature β, represent a fraction $\chi_i(\beta) = \omega_i \exp(-E_i\beta)/Z(\beta)$ of the posterior (Boltzmann distributed) space. The partition function is estimated by numerical integration as

$$Z(\beta) \approx \sum_i \omega_i \exp(-E_i \beta).$$

The expected value of any thermodynamic observable, $Q(\Omega)$, at any temperature can also be estimated as

$$\mathbb{E}(Q|\beta) \approx \sum_i \chi_i(\beta) Q(\Omega_i)$$

The algorithm terminates when a specific criterion is reached, for example, the estimate for $Z(\beta)$ has converged for the temperature of interest. Other criteria may be more appropriate for some systems (39). Algorithm 1 summarizes the nested sampling algorithm.

Algorithm 1 The Nested Sampling Algorithm

```
1) Sample $K$ conformations uniformly (w.r.t. conformation space), $\{\Omega_1, \Omega_2, \dots \Omega_K\}$ and calculate their
energies $\{E_{\Omega_1}, E_{\Omega_2}, \dots E_{\Omega_K}\}$.
2) Remove the conformation with highest energy from the active set; save it as $(E_1, X_1, \Omega_1)$.
3) Generate a new conformation $\Omega$, sampled uniformly from the conformations with energy below $E_1$,
and add it to the active set.
4) Repeat Steps 2 and 3, generating $(E_2, X_2, \Omega_2), (E_3, X_3, \Omega_3), \dots$.
```

Sampling techniques such as parallel tempering and simulated annealing struggle to sample systems which undergo first order transitions, such as the system shown in Figure 1.2. Starting at diagram E, SA has to find the volume of phase space with energy U_x which has non-negligible probability mass. This volume is exponentially small in comparison to the phase space of conformations with energy U_Y and in some cases it can be effectively impossible to find it. Furthermore, going from diagram D to A, the proportion of samples from each Gaussian needs to be kept in equilibrium. Added to this, the fact that the temperature of the phase transition is not known *a priori* means the temperature schedule may miss the transitions entirely. Similar concerns exist for PT. There may be no replicas with temperature near T_c, and even if there are, for example if an adaptive temperature schedule is used, it can still be very hard to equilibrate.

Phase transitions, however, are not a problem for nested sampling, which takes steps equidistant in log phase space volume, taking smaller steps (in energy) when the space shrinks more quickly. It just marches down from U_Y to U_X, unaware that there was a problem for thermal methods.

In this section we have considered a variety of sampling algorithms without discussing their application to research problems. In §1.3.6 and §1.3.7 we discuss the application of these algorithms to the study of proteins and the results they have produced. For more details of general atomistic sampling, however, we refer the reader to (10) and references therein.

1.3 Proteins: Modelling and Applications

1.3.1 Overview

Proteins are essential to life. Some are enzymes, catalysing cellular chemical reactions. Others are antibodies, recognising pathogens. Yet more control transport into and out of cells or sense and act on signals generated by the organism or environment. Regulatory proteins regulate and control cellular processes such as gene transcription; structural proteins, such as keratin and actin, form tissue such as skin and muscles; and motor proteins convert chemical energy to mechanical energy, for example in muscle contraction. The incorrect behaviour of proteins is implicated in a large range of disease such as Alzheimer's or Type II diabetes. The intricate three-dimensional structures of proteins are finely tuned, through evolution, to perform their highly specialised functions. The diverse structures proteins adopt explain the large variety of functions they are able to perform.

Protein science, the study of these molecules, is a very wide field, encompassing theory and techniques from chemistry, molecular and cellular biology, evolutionary biology, bioinformatics, computer science and biophysics (44). Over the last 70 years proteins have been studied both experimentally, using techniques including X-ray crystallography (45, 46), NMR spectroscopy (47), circular dichroism (48) and fluorescence resonance energy transfer (FRET) (49); and computationally, from simple lattice models (50) through to full all-atom models with more realistic force fields (51).

In this work we mainly focus on globular (water-soluble) proteins. However, there is another important class of proteins: membrane proteins. Membrane proteins are found in water-deficient environments and hence are subject to different conditions to globular proteins and so behave differently. For example, exposed charges, whilst often favourable in water, are severely energetically disadvantageous when in contact with a hydrophobic ('water-hating') membrane.

The following sections describe the structure of proteins, the theory of protein thermodynamics, the different types of protein models used to study proteins computationally, how the parameters of these models are inferred and how conformations are sampled using them. Finally, there is a summary of applications of protein models. However, discussion of an important application, that of protein structure prediction, is deferred until later.

1.3.2 Protein Sequence and Structure

The building blocks of proteins are amino acids and there are 20 standard, *proteinogenic* amino acids. With the exception of one, proline, all have a standard molecular structure (Figure 1.3 (left)) shown with the standard atom names), differing only in their side chain. For example, glycine (code G) has a single hydrogen side chain, whereas serine (S) has a side chain CH_2-OH. The side chain of proline (P) covalently bonds to both the C_α and N atoms; see Figure 1.4. Different side chains endow the amino acids with different properties. For example, some amino acids are hydrophobic ('water-hating') and others are hydrophilic ('water-loving').

A protein is a linear sequence of amino acids joined by peptide bonds (Figure 1.3 (right)) forming a protein chain. Short protein chains are called peptides. Proteins are defined by a specific sequence of amino acids (also called residues). An organism's protein sequences are encoded within its DNA. The repeating -(N-C_α-C')- pattern denotes the protein backbone and the two ends of the protein are called the N and C termini.

Over 50 years ago, the chemist Linderstrøm-Lang described a hierarchy of protein structure (53), the primary, secondary and tertiary structure of a protein; these terms are defined below. A further level, the quaternary structure, was later added to describe how separate protein chains assemble into larger structures.

The primary structure, or sequence, of a protein is precisely an ordered list of its constituent amino acids. Due to the polar nature of the N-H and C-O covalent bonds (H is slightly positively charged, O negatively so), an electrostatic 'hydrogen bond' can form between water and a residue or between two different residues. Certain intra-protein hydrogen bond patterns are energetically favourable and hence common motifs, α-helices and β-strands, are found in most proteins. The specific pattern of helices,

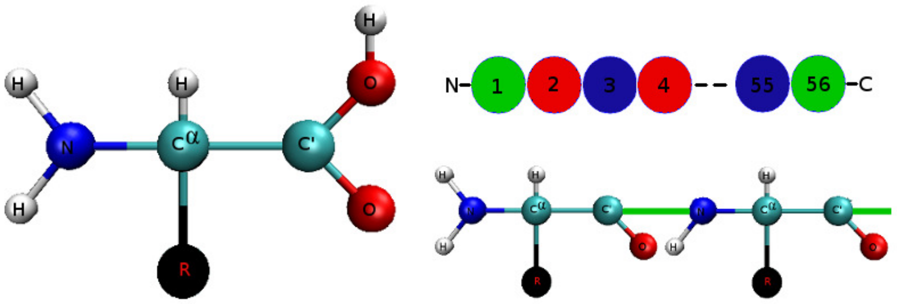


Figure 1.3: Left: The molecular structure of an amino acid. The atoms are labelled with their standard names and R represents the amino acid side chain. Right Top: A protein is a linear sequence of amino acids and the ends of the protein are called the *N* and *C* termini. Right Bottom: The first two residues of a protein joined by a peptide bond (in green) between C′ of residue 1 and *N* of residue 2. All atomic models have been created using VMD (52).

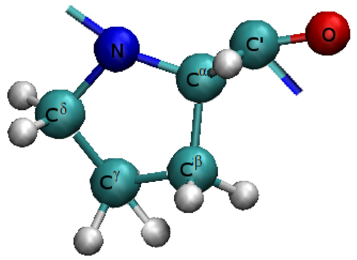


Figure 1.4: Unlike the sidechains of the other amino acids, which are only bonded to the C_α atom, the side chain of proline, as shown, covalently bonds to both C_α and N atoms. Therefore, proline is technically an imino rather than amino acid. In amino acids the side chain atoms are labelled with the β-atom bonded to the backbone C_α atom, the γ-atom bonded to β-atom etc. The C_β, C_γ and C_δ atoms of proline are labelled.

strands and coil (other) residues is the secondary structure of a protein.

In an α-helix, the hydrogen atom from residue i bonds with the oxygen atom of residue $i-4$; continuing this pattern leads to a helix with 3.6 residues per turn, as shown by Figure 1.5 (left). Other helices, whilst significantly rarer, are also found in proteins, specifically, 3_{10}-helices, where the bonding pattern is $i \to i-3$; and π-helices, which have bonding $i \to i-5$. A β-strand is typically a sequence of 5-10 residues hydrogen bonded with a set of consecutive residues elsewhere in the protein. β-strands can hydrogen bond in parallel or anti-parallel directions and when bonded form β-sheets; see Figure 1.5 (right).

The reason α-helices and β sheets are energetically favourable can be understood in terms of backbone dihedral angles ϕ and ψ and the Ramachandran plot. The dihedral angle of a chain of atoms a_1, a_2, a_3, a_4 is the angle between the planes containing a_1, a_2, a_3 and a_2, a_3, a_4. In a protein, ϕ_i is defined as the dihedral angle between C′-N-C_α-C′, ψ_i between N-C_α-C′-N and ω_i between C_α-C′-N-C_α, where i runs over the residues of the protein. See Figure 1.6 (left). Due to the nature of peptide bonds, the four atoms $\{C_\alpha, C', N, C_\alpha\}$ are always close to planar and hence ω is always close to 180° (*trans* conformation) or 0° (*cis* conformation).

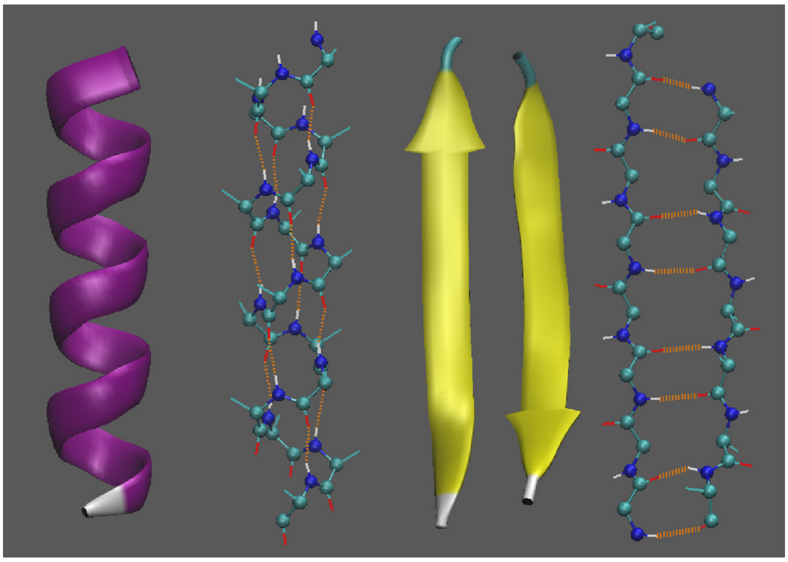


Figure 1.5: Left: A cartoon picture showing an α-helix and the corresponding full atom representation. Right: A cartoon picture showing a pair of anti-parallel β-strands with the corresponding full atom representation. Only main chain and C_β atoms are shown. The backbone atoms are highlighted and the hydrogen bonds are coloured orange.

With the exception of proline residues, the *cis* conformation is extremely rare due to the energetically unfavourable steric repulsion of consecutive C_α atoms when compared to the *trans* case in which the C_α atom is close to the much smaller amide hydrogen. The special nature of the proline sidechain makes the *cis* conformation much less unfavourable, as in the *trans* case the small hydrogen atom is now replaced by the massive C_δ atom (see Figure 1.4). Around 10% of peptide bonds preceding proline residues are found in the *cis* conformation.

Unlike ω, ϕ and ψ are allowed to vary and provide most of the conformational freedom of the protein main chain. A Ramachandran plot is a graph showing the dihedral angles of a protein (Figure 1.6 (right)). Only certain regions of the graph can be occupied, as dihedral angles from other regions would lead to steric clashes between backbone or C_β (the carbon atom in the side chain bonded to C_α) atoms. Both the α-helix and β-strand have standard dihedral angles which are in the sterically allowed regions of the graph. Both proline, with its special sidechain, and glycine, which lacks C_β and hence has more freedom, have different allowed regions to other amino acids. Certain local motifs of secondary structure elements, such as a β-hairpin (two anti-parallel β-strands connected by a short turn) or a β-α-β unit (two interacting parallel β-strands separated in sequence by a single helix), are frequently found in proteins, and these motifs are often called the 'supersecondary' structure.

The tertiary structure of a protein is how the secondary structure elements combine to form the full three dimensional structure. The tertiary structure is determined by many factors. These include the hydrophobic effect; non-polar side chains being buried in the core of the protein away from the solvent (54, 55); electrostatic charges, both involving hydrogen bonds and salt bridges between charged amino acids (56); dense side chain van der Waals packing (57); the existence of metal ions (58) and disulphide bridges, where the sulphur atoms of two cysteine (C) amino acids form a covalent bond (59).

Typically large single chain proteins are composed of domains, single subunits capable of folding independently of the rest of the chain, connected by somewhat disordered loops. Certain domain folds are particularly common, such as the three helix bundle. For the last 20 years, domain structures have

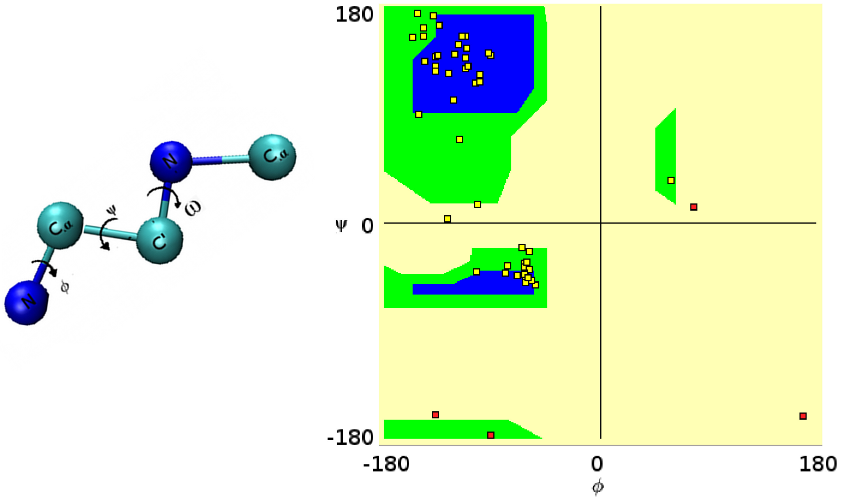


Figure 1.6: Left: The dihedral angles ϕ, ψ and ω. Right: A Ramachandran plot showing the dihedral angles of protein G (Figure 1.11 (left)). The shaded blue regions are the most stable; where ψ is positive this indicates a β-strand and when it is negative, an α-helix. The shaded green regions are allowed but more strained. The yellow areas are less favourable due to steric clashes and are significantly less common. Glycine, with its smaller sidechain, has more freedom than other amino acids, and the glycine residues of protein G have been highlighted in red.

been classified and databases of them maintained (60, 61).

In this work we focus on the large number of proteins which have well-defined tertiary structure essential for their function. Since the pioneering work of Anfinsen and colleagues (62), it has been understood that proteins (at least small, globular proteins) fold into their tertiary (native) structure reversibly, leading to the conclusion that the structure of a protein is fully encoded in its primary sequence and that the native state is thermodynamically stable. However, recently, there has been a lot of interest in intrinsically disordered proteins, that is proteins which do not adopt a well defined structure, until, say, they bind to a ligand; they are now believed to be significantly more numerous and important than originally thought (63, 64).

1.3.3 Protein Thermodynamics

In 1969 the molecular biologist Cyrus Levinthal asked the question: how can a protein fold into its native state in the millisecond to second timescales that are observed? For simplicity, assume each residue can take two possible states, α-helical or β-sheet. Then a 100 residue protein would have $\approx 10^{30}$ possible conformations, and assuming an interconversion rate of $\sim 10^{-13}s$, it would take of the order of 10^{10} years for the peptide to explore its conformational space. This argument, named the Levinthal paradox, implies that protein folding is not a simple diffusive process, and this has lead to the idea of a protein folding funnel (65–68).

Protein Folding Funnel

The free energy landscape of a system behaving according to Levinthal's diffusive process is shown by Figure 1.7 (left). With the exception of the native state, the landscape is flat and hence a random search is required in order to 'stumble upon' the native state. In contrast, Figure 1.7 (right) shows a protein folding funnel. The protein starts from one of a large number of conformations with high energy and hence has high entropy. The chain decreases in energy towards the native state, a state with low entropy.

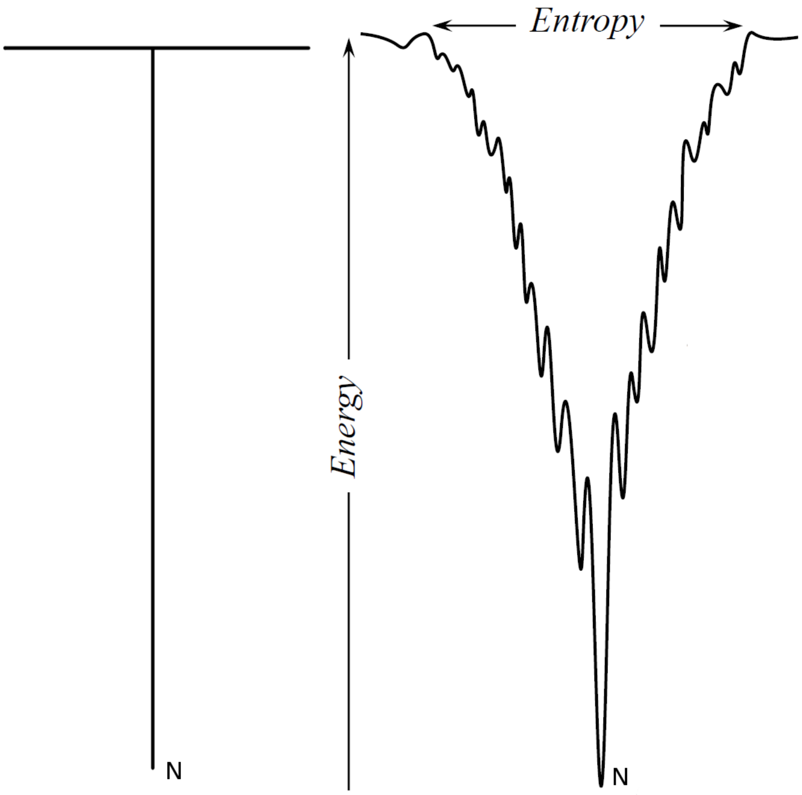


Figure 1.7: Left: The free energy landscape of a system behaving according to Levinthal's diffusive process. With the exception of the native state, the landscape is flat and hence a random search is required in order to 'stumble upon' the native state (N). Right: A protein folding funnel. The protein starts from one of a large number of conformations with high energy and entropy. The chain decreases in energy towards the native state (N), a state with low energy and entropy.

A realistic folding funnel of a naturally occurring protein has rugged sides corresponding to metastable states or kinetic traps which slow the steady march of the protein toward its native state. The free energy difference between the native state and other free energy minima is of the order of 10 kcal/mol (69), which equates to only a couple of hydrogen bonds. This implies that naturally occurring proteins are only marginally stable, which has been shown to be evolutionarily advantageous (69). In contrast, randomly chosen heteropolymers typically have very rugged energy landscapes with no clear native state: the free energy gap between the global minima and other minima is small and the dynamics are glassy (70).

One theory of protein folding is the hierarchical folding theory of Baldwin and Rose (71, 72). According to this theory, a protein folds by first forming local secondary structural elements such as α-helices and β-strands, and then these structures interact with each other to form the native structure.

Initially the unfolded protein collapses from an extended chain into a 'molten globule' state. At this time the protein is relatively compact, has a substantial amount of correct secondary structure and yet has more flexibility and looser hydrophobic packing compared to the native state. In order to form the native state, the secondary structure elements come together, forming a tighter hydrophobic core with the amino acid side chains interlocking tightly.

In order for the native state to form, the protein has to overcome free energy barriers (66) in a similar way to chemical reactions requiring an activation energy in order for the reaction to proceed. The free energy barriers correspond to regions of increasing free energy on the free energy landscape, which must be traversed in order to reach the native state. The conformations at these local maxima are called transition states.

From an experimental point of view, it is not possible to visualise the high dimensional free energy landscape. However, protein folding pathways can be investigated through a procedure known as Φ-analysis (73). Φ-analysis allows investigation of the transition state, which is by definition unstable and short lived and therefore not amenable to standard structural methods.

Figure 1.8 (left) shows a reaction diagram for a non mutated (wild-type) protein. Although the free energy of the native state is lower than that of the unfolded state, in order for the protein to fold, activation energy is required in order to reach the transition state. If a residue is making the same interactions in both the native state and the transition state then, when it is mutated, both the transition and folded states are destabilized by similar amounts, as shown in the reaction diagram Figure 1.8 (middle). However, if the residue is not making the same interactions, then only the folded state is destabilized and the reaction diagram Figure 1.8 (right) is seen. The ratio of the destabilization of the transition state to the destabilization of the native state, $\Phi = \Delta\Delta G^{\ddagger}/\Delta\Delta G$, can be measured for different residues by estimating rate constants, and these results can be used to build up a picture of the transition state.

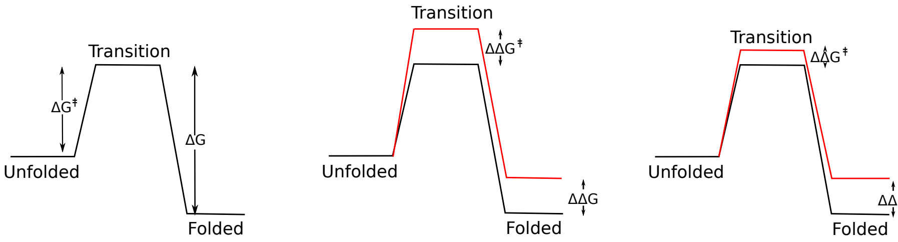


Figure 1.8: Left: A reaction diagram for a non-mutated (wild-type) protein, where a free energy barrier of height $\Delta G^{\ddagger}$ must be overcome in order for the protein to fold. Middle: The reaction diagram of the protein when a residue which has the same interactions in both the native state and the transition state has been mutated (red). The transition and folded states are destabilized by similar amounts and $\Phi = \Delta\Delta G^{\ddagger}/\Delta\Delta G$ is close to 1. Right: The reaction diagram of the protein when a residue which does not have the same interactions in both the native state and the transition state has been mutated (red). The folded state is destabilized significantly more than the transition state and Φ is close to 0.

The situation described by Figure 1.8 (left) is the most simple case. Many proteins have more complicated reaction diagrams, including meta-stable intermediate states, corresponding to local minima of the free energy landscape, and even multiple folding pathways, whereby a different folding rate is attained depending on which folding pathway is taken.

Finally, from a statistical thermodynamics viewpoint, there are theoretical arguments suggesting that denaturation of globular proteins is a (quasi) first order phase transition (74); experimental evidence such as circular dichroism spectra (75) and experimental heat capacity curves (76) also supports the conclusion that the transition is all-or-none. This first order transition has implications for the sampling of protein conformations *in silico*. The nature of this transition explains the free energy barriers and activation energy required for a protein to fold. For further discussions concerning protein thermodynamics we refer the reader to (77).

It is very useful to form general theories about protein folding and thermodynamics. However, it is important to note that the complex balance of competing forces can produce very different behaviours in different proteins and that these sequence-specific details are of crucial importance. For example, ubiquitin is a 76 residue protein with a folding rate of $\sim 3ms$, whereas the 73 residue protein α3D folds around three orders of magnitude faster at $\sim 3\mu s$ (78). Equally noteworthy, Karanicolas and Brooks have shown that two proteins with the same native fold follow very different folding pathways, with protein G initially folding its C terminal hairpin, in contrast to protein L which first folds its N terminal hairpin (79).

1.3.4 Protein Models

In the 1950s, Linus Pauling and Robert Corey used simple protein models to predict the existence of α-helices and β-sheets before they were discovered experimentally (80). Since that time protein models have continued to improve our understanding of protein structure, stability and thermodynamics. Early models, such as the HP lattice model of Lau and Dill (50), gave an insight into the general principles of protein folding, rather than protein-specific details.

In the HP model, each residue is either hydrophobic (H) or polar (P) and is represented by a single bead positioned on a square lattice; placing two hydrophobic residues on adjacent sites is energetically favourable. This model has been used to study hydrophobic collapse and, for small proteins, has the advantage of being simple enough to exhaustively explore the *entire* conformational space. With increases in computational power, more complicated and realistic models have been developed and in this section we give an overview of the types of models which are currently in use.

Protein models with vastly different levels of complexity and accuracy have been developed. The most accurate, and hence computationally expensive, are QM/MM approaches which combine molecular mechanics force fields with regions captured in full quantum mechanical accuracy. Due to the vast expense of using these models, they are typically only used when covalent bonds are being formed or broken, as in this case *only* QM approaches can be used; see the recent review on the use of QM/MM methods with proteins and biomolecules (81).

More typically, models of proteins can be developed from standard all-atom (AA) force fields, such as AMBER (82) or CHARMM (83). These models, often including explicit water molecules, have been used in simulations of proteins unfolding (84, 85) and even the folding simulations of small proteins (51). AA models provide a high level of accuracy as they typically use complex energy functions in order to capture all pairwise atomic interactions. These energy functions are computationally expensive and so AA models cannot be used to capture the behaviour of large proteins or the long time behaviour of small systems.

In order to study larger systems for longer timescales, coarse-grained (CG) protein models have been developed. CG models are simpler than AA models, for example representing the whole side chain with a single bead, yet complex enough to still capture the physics of interest. CG models can be divided into two broad classes: structure independent, where no prior knowledge of the structure is required, and structure dependent, where the model is built from knowledge of the structure of the system (86, 87).

Gō models (88) are examples of structure dependent CG models. They assume that only native interactions contribute to the overall shape of the folding energy landscape and an attractive potential is applied to residues in contact in the protein native state. These models have been used to investigate the folding landscape (funnel) of proteins (89) and protein unfolding due to the application of force (90).

Elastic network models (ELN) are another type of structure dependent model (91–93). In these models harmonic springs are attached to atoms (or beads) which are close in the native structure. ELN models allow the study of folded protein dynamics, for example by normal mode analysis, that is projecting the network along its eigenvectors (94), or rigidity analysis, which determines flexible regions of proteins in order to determine plausible, large-scale, conformational changes (95).

In contrast, structure independent models typically rely on the physical and chemical properties of constituent atoms of the protein and solvent and are used for the study of protein aggregates (96) or when native structure is unknown (in the case of structure prediction) or intrinsically disordered (97). CG structurally independent models have not been as useful as those which incorporate structural information because in proteins very specific side chain interactions contribute to protein stability and in CG models these interactions are not usually modelled to a high enough level of detail. Due to this, there has been a lot of work developing models which sit somewhere between the two paradigms, so called weakly-biased models, where a small amount of structural information, often secondary structure information, is incorporated into a structural independent model. Examples include MARTINI (98) and AMH (99) which can be tuned to balance the structural bias and physical force field.

As an example of a weakly-biased CG protein model, we describe CRANKITE (100, 101), a protein model designed to stabilize secondary structure elements in room-temperature simulations.[6] CRANKITE includes all heavy main chain atoms, together with the amide hydrogen and side chains represented by C_β atoms; see Figure 1.9. There are 3 degrees of freedom per residue, the dihedral angles ϕ and ψ and τ, the C_α valence angle. All other valence angles, bond lengths and dihedrals are fixed, notably the peptide bond which is exactly planar. As is common for CG models, the effect of the solvent is modelled implicitly through the energy function rather than the explicit modelling of water molecules.

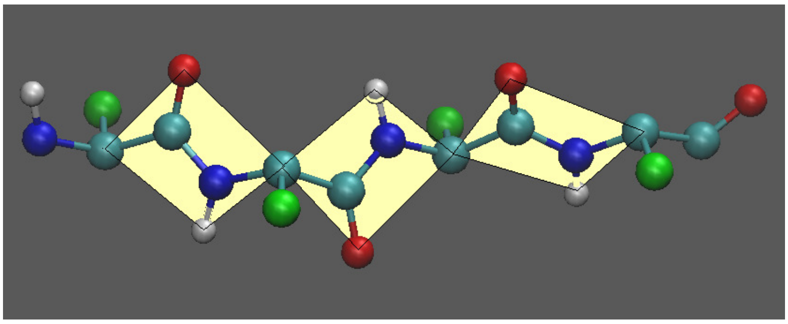


Figure 1.9: The CRANKITE protein model. All heavy main chain atoms and amide hydrogens are modelled. Side chains are represented by C_β atoms (shown in green). The peptide bond is exactly planar (as shown by yellow planes) and the dihedrals ϕ and ψ as well as τ, the C_α valence angle, are free to vary.

A wide variety of CG models have been proposed, ranging from a very high level of coarse-graining, such as a single C_α (102, 103) or C_β (104) bead per residue, to a much more detailed representation, for example, representing sidechains by up to 6 beads (98).

The energy functions used by protein models attempt to capture the various factors involved in protein folding, stability and dynamics. AA models typically use standard energy functions as described in §1.2. On the other hand, energy functions for CG models of varying levels of complexity and expense have been developed, aiming to balance the need for accuracy with computational expense. Once again we provide an example by focussing on the energy function of CRANKITE. Given a protein sequence **R**

[6] We describe CRANKITE as found in (101). See Chapters 2 and 3 for recent improvements to the model.

of length N and a specific conformation Ω the energy $E_{\mathbf{R}}(\Omega)$ is given by

$$E_{\mathbf{R}}(\Omega) = \sum_{i=1}^{N} E_i^{\mathrm{B}} + \sum_{i=1}^{N}\sum_{j=1}^{i}(E_{ij}^{\mathrm{vdW}} + E_{ij}^{\mathrm{HB}} + E_{ij}^{\mathrm{SC}}).$$

E_i^{B} is a harmonic function keeping τ restrained and E_{ij}^{vdW} is a simple model of the van der Waals atomic repulsions. How well simplified van der Waals functions capture the intricate side chain interactions and packing when compared to more sophisticated Lennard-Jones functional forms is an interesting research question and is discussed in Chapter 3. E_{ij}^{HB} is a hydrogen bond term capturing the essential electrostatics required for secondary structure formation; see (105). In AA models this term is not usually required as H-bonds are implicitly incorporated into a more general electrostatic function. Finally E_{ij}^{SC} provides the weak-bias part of the model and is the only energy term which depends on $\mathbf{R}$. It forces α-helix formation for residues known to form helices, and for residues known to form β sheets, it forces local extending of the chain to form strands and also pulls together interacting β-strands. Often CG models will include a hydrophobic energy term (two hydrophobic residues in contact is energetically favourable) which is designed to allow a hydrophobic core to form. This effect is actually caused by surrounding water molecules but these are not explicitly modelled.

When developing CG models, as well as the trade off between computational accuracy and expense, it is important to consider the transferability of the model to other, often similar, systems. Transferability of CG models is an active area of research (106, 107).

1.3.5 Parameter Inference

Given a protein model and energy function, it is a non-trivial task to optimize parameters, such as hydrogen bond strength or atomic radii, to produce accurate and reliable results. A wide variety of techniques have been developed. General purpose AA models optimize their parameters to match quantum mechanical calculations (108) or experimentally derived properties of small molecules (109).

In contrast, when optimizing parameters for CG models, especially those which include some form of bias, some terms of the energy function may not have a well-defined physical force associated with them. In these cases force field parameters should be optimized so that the native state is found at the (physiologically relevant temperature) global free energy minimum (62).

Traditionally CG models have used statistical-knowledge-based potentials (110) and the parameters have been tuned to best reproduce features, such as dihedral angle distributions or atomic distance distributions derived from a training set. This procedure relies on the *Boltzmann hypothesis*, the assumption that within native structures the features are statistically independent and are distributed according to the Boltzmann distribution. There is some empirical evidence for this hypothesis (111), yet the statistical independence of features is likely to be a poor assumption.

Over twenty years ago Maiorov *et al.* and Goldstein *et al.* developed native structure discriminant methods for parameter inference (112, 113) and these remain popular to this day (114). These methods optimize the parameters so that the native state has the lowest energy when compared to a decoy set of protein-like conformations. A disadvantage of these methods, however, is that they do not take temperature and hence protein thermodynamics into account: only the strength of the intermolecular forces relative to the decoys is used in the parameter estimation.

An alternative class of optimization algorithms use the principle of maximum likelihood (ML). Given a data set of observed experimental (or computationally generated) samples $\mathbf{\Omega}$ and a set of force field parameters $\mathbf{\Theta}$, an appropriate likelihood function, typically the Boltzmann distribution at an appropriate temperature, $L(\mathbf{\Theta}|\mathbf{\Omega})$, is introduced. ML methods tune their force field parameters in order to maximize this likelihood function, iteratively improving by following the gradient of the logarithm of the likelihood function.

Models which use parameters that maximize the likelihood function will produce the (suitably defined) *closest* distribution to the original dataset and ML (also called relative entropy) approaches have been used successfully to infer parameters for CG water (107) and polyalanine (115) models, using samples from AA models as the data set. Parameter estimation methods for general protein models using the PDB database as the data set have also been developed (105, 106, 116).

Contrastive Divergence

For the simple case of a single parameter $\mathbf{\Theta} = \{\theta\}$, a single conformation $\mathbf{\Omega} = \{\Omega_0\}$ and setting the inverse thermodynamic temperature $\beta = 1$, the gradient of the log-likelihood required for ML methods is given by

$$\frac{\partial \ln L}{\partial \theta} = \left\langle \frac{\partial E(\Omega, \theta)}{\partial \theta} \right\rangle - \frac{\partial E(\Omega_0, \theta)}{\partial \theta},$$

where $E(., \theta)$ is the (potential) energy function using parameter value θ and the angular brackets correspond to the thermodynamic expectation of the system using energy function $E(., \theta)$. A full derivation of this and the general case can be found in Chapter 3.

Although $\partial E(\Omega_0, \theta)/\partial\theta$ can be calculated directly, the thermodynamic average can only be accurately estimated by running an MC or MD sampling algorithm until equilibrium is reached and then taking the expectation of a large number of equilibrated samples. This is an expensive procedure which needs to be carried out for a different θ for every iteration of the ML procedure. For example, Winther and Krogh estimate the thermodynamic average by running extensive REMD simulations for each ML iteration (106).

A few methods which aim to reduce the computational expense have been developed. For example Shell *et al.* reweight samples from one iteration for use at later iterations so as to reduce the number of long equilibration runs required (115). Here we focus on an alternative method, known as contrastive divergence (CD). CD is a statistical machine learning technique, initially developed to efficiently learn the parameters of Boltzmann machines (117).

For each ML iteration of the CD procedure, rather than running until equilibration, Ω_0 is evolved only K MC steps to conformation Ω_K. K is a tunable parameter and theoretically can be as low as 1. Rather than using the true log-likelihood gradient we replace it by

$$\frac{\partial E(\Omega_K, \theta)}{\partial \theta} - \frac{\partial E(\Omega_0, \theta)}{\partial \theta}$$

when updating θ.

The idea behind this approximation is that even after only K steps, the data distribution has drifted towards the equilibrium distribution; Ω_K is closer than Ω_0 to the equilibrium distribution (for the current value of θ), where *closer* is appropriately defined. The drift in the observed energy gradient can then be used to guide the update procedure. A full justification and further discussion can be found in Chapter 3.

CD is significantly computationally cheaper than traditional ML methods and therefore a larger data set can be used for parameter inference. For example, Winter and Krogh were restricted to 24 different 11–14 residue-long protein fragments (106), whereas using CD, Podtelezhnikov *et al.* were able to use a database of 247 protein PDB files as a data set (105). A large data set is important for transferability; Winter and Krogh found their force field performed poorly when used with proteins and peptides not in their data set.

1.3.6 Conformational Sampling

In order to use a protein model a simulation needs to be run, generating conformational samples which can then be used in analysis. If kinetic parameters such as accurate folding or transition rates are being studied then MD simulations should be run, typically keeping temperature and pressure constant; care must be taken to use a thermostat which preserves these properties.

If dynamic properties of the system, such as folding rates, are not required, then alternatives to MD can be used to generate samples. For example, MC sampling is effective for CG models where the only degrees of freedom are the dihedral angles ϕ and ψ. An example of a proposal move is a crankshaft rotation; the axis between two C_α atoms is chosen and the residues between these two atoms are rotated around this axis, as in Figure 1.10. C_α atoms close in the chain are selected so that the MC moves are local. In order for τ, the C_α valence angle, to remain fixed, only certain combinations of rotations are allowed and their calculation, the 'loop-closure' problem, is computationally expensive (118). CRANKITE, inspired by the study of large-scale DNA properties, allows τ to vary, thus negating the need for a computational solution to the loop-closure problem (100).

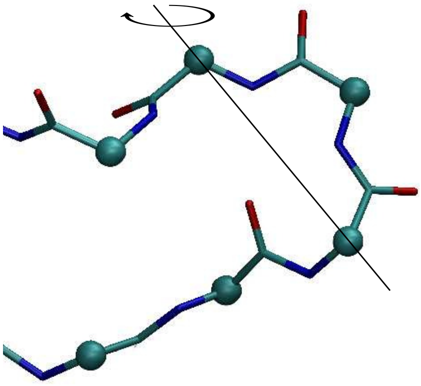


Figure 1.10: An example crankshaft rotation. The axis of rotation is chosen as the axis connecting two C_α atoms (shown as spheres). The portion of the chain between these atoms (to the right in this figure) is rotated; the rest of the chain is fixed. C_α atoms close in the chain are selected so that the MC moves are local.

Compared to AA models, CG models are able to explore the phase space much more easily, not only because energy evaluation is quicker but also because high frequency vibrations (such as those of covalently bonded hydrogen atoms) have been 'coarse-grained' away so that larger timesteps or MC moves can be taken. Another advantage of CG models with regards to exploration is that they smooth the energy landscape, removing some of the ruggedness and local traps which slow down the exploration of AA models. This is particularly noticeable with Gō models, where only native interactions are included, thus smoothing the landscape enormously (87). Also, for CG models, statistical averages can more easily be calculated as they, unlike AA models, are often cheap enough to run large numbers of simulations in order to combine their results.

Alongside MD and MC, the advanced sampling techniques described in §1.2.3 have been applied to sampling protein conformations. REMD is a very common choice when thermodynamic data is required because exchanges between low and high temperature replicas ensure chains do not become trapped in basins. Also, researchers are often interested in the change of behaviour as a function of temperature, so the samples from higher temperature replicas are of value in their own right. For AA simulations, the crystal structure is typically used as a starting condition for each replicate to save the computational expense of actually folding the protein or in order to conduct unfolding–refolding simulations. However, REMD has been used to fold small mini-proteins such as the 20 residue Trp-cage (119).

As REMD has been so successful for sampling protein thermodynamics, there have been algorithms proposed which use REMD to obtain protein kinetics (120, 121) and even transition path analysis (122). Biased MD algorithms such as umbrella sampling (34) and self-guided MD (123) have also been used to generate conformations, aiming to explore as much of the phase space for as little expense as possible.

The exploration of protein potential energy surfaces (PES) has generated specialised sampling algorithms, such as David Wales' discrete path sampling method (124, 125), which has been applied to small proteins and peptides in order to explore PESs. Discrete path sampling proceeds by building a database of potential energy minima and transition states, clustering the minima and using master equation dynamics to estimate the kinetic parameters of the underlying system.

Finally, further examples of sampling algorithms such as discrete MD (96, 126, 127) and normal mode projection (94, 128) have also been applied to proteins. Discrete MD is an event-driven variant of MD, where the collision dynamics of systems interacting with discontinuous potentials can be solved exactly and normal mode dynamics projects the system along the eigenvectors of the mass-corrected force constant matrix, allowing the study of large scale conformational changes.

1.3.7 Applications

We end this section with a discussion of protein model applications and the results derived from their study. A final application, that of protein structure prediction, is deferred until later.

From the earliest lattice models, computational simulation has been used to study the process of protein folding, the characterization of folding pathways and the dynamics and thermodynamics of globular proteins.

Recent work by the Shaw group has shown the viability of using MD with AA models to fold small proteins and probe the protein folding pathways to atomic level accuracy (51), although special purpose hardware was required. However, it is important to remember that AA protein force fields have been optimized for proteins in their native state and it is unknown how well they model unfolded chains (129). For example, researchers have studied the accuracy of AA force fields when modelling small peptides such as alanine dipeptide (130), trialanine (131) and the five residue peptide Met-Enkephalin (132); they are small systems whose peptide bonds are believed to be behave reasonably similarly to those found in unfolded proteins. Results have been compared with other AA force fields, QM calculations or experimental work with mixed results; good agreement between different methods and force fields is often found (131) but this is not always the case (130).

Alongside protein folding studies, researchers have used protein models to study the behaviour of proteins unfolding. For example, Dudko *et al.* have used a Gō model to study the behaviour of proteins unfolding by the application of force (90). They show how the choice of reaction co-ordinate is crucial in understanding the energetic barriers to unfolding. These simulations can be directly compared to the experiments where single proteins are pulled apart by the application of force (133). Protein unfolding is a much faster process compared to folding and hence is also accessible to AA models. For example Li and Daggett characterized the (unfolding) transition state of Chymotrypsin inhibitor 2 (84), and their later work compared simulations and experiments of the unfolding of the Engrailed Homeodomain protein (85).

The thermodynamics of proteins has also been extensively studied with protein models. The heat capacity of proteins can be experimentally measured and so its calculation *in silico* is highly desirable, although very challenging to compute. For example, Yeh *et al.* calculate the heat capacity for an SH3 domain, starting the simulation from the crystal structure (134), and Lee and Olson calculate the heat capacity for the Trp-cage (135). The position of the peak in the heat capacity curve corresponds to the temperature at which the protein unfolds, and alternatives to the heat capacity, such as the proportion of native contacts (Q) and even the radius of gyration R_g, are often reported. It is also common to present free energy landscapes projected onto suitably chosen reaction co-ordinates. For example, Shea *et al.* present the free energy surface as a function of Q and R_g for an SH3 domain (136) and Zhou studies the effect of explicit and implicit water on the free energy surface of a β-hairpin using R_g and number of H-bonds as the reaction co-ordinates (137).

The potential energy landscapes and folding funnels of proteins have also been studied and visualised using disconnectivity and later scaled disconnectivity graphs (138, 139). Koga and Takada have used protein models to perform *in silico* mutation analysis, mimicking experiments in order to study the mechanism of the rotary motor F1-ATPase (140).

Protein models have also been used to study peptide aggregation, one of the processes known to be heavily involved in diseases such as Alzheimer's. Nguyen and Hall studied the sensitivity of fibrillization on temperature and peptide concentration (96). Their simulations provided evidence for the nucleated fibrillization hypothesis; an ordered nucleus is formed from a small amorphous aggregate and this is then followed by rapid fibril formation. Fawzi and coworkers use a CG model of the Alzheimer's $A\beta_{1-40}$ peptide in order to study the propensity of different protofibiril seeds to form full fibirils, their patterns of growth and level of stability (141).

Since the 1970s there has been success using protein models to study protein-protein docking (142), that is building models of proteins known to interact and using the models to predict the interaction site and relative orientations of the molecules. Initial models treated the molecules as rigid bodies, but with the increase in computational power, more recent models allow for flexible docking. There has also been success combining models with experimental results, improving our understanding of protein docking (143). See the special review issue of *Proteins: Structure Function and Bioinformatics* (144) for details of the recent progress in this field.

Another application of protein models is in protein design. Using computational techniques, Kuhlman *et al.* successfully designed a 93 residue protein, Top7, which was shown experimentally to be stable and have the designed tertiary structure (145). The Mayo lab have been at the forefront of computational protein design (146, 147) and they have incorporated protein models and computational design into directed evolution experimental pipelines. Directed evolution is an experimental approach to designing proteins. For example, to improve binding affinity to a specific ligand, a library of sequences is taken and random mutations are performed, and the sequences with the highest affinity are kept for the next iteration. Protein models can be used as a filter for desirable sequences before *in vitro* experiments (148).

Finally, one of the main criticisms of CG models is that they may not be sufficiently accurate and therefore that the conclusions drawn from the models may not be relevant to real systems. Therefore, it is important to compare results to experiment (97, 149), or failing that, all-atom models of the same system (107, 150). However, if this comparison is not always shown or is very general, the results can only provide insights into possible or qualitative behaviours of the system. Due to the increasing amount of experimental data, the increase in computing resources and the maturing nature of the field, recent CG modelling work is, generally, more able to compare to experimental or AA work than it was in the past.

In recent years, CG models have improved our understanding of biophysical systems. As computational power increases, larger systems will be able to be studied using AA models (51), and for CG models to maintain their utility, new models will have to be developed which study larger, even mesoscale, systems (151, 152). Further work is also required to improve the transferability of CG models and in inferring their parameters. Novel ways of using CG models, for example by running hybrid CG-AA models (153), may also be an interesting line of enquiry.

1.4 Protein Structure Prediction

1.4.1 Overview

The prediction of the tertiary structure of proteins given their primary sequence is one of the central problems in computational biology. Beginning in the 1960s, with the prediction of α-helices from residue identity (154), the problem remains an active area of research to this day (155, 156). Protein structures are required in order to determine the mechanism of action of a protein or to use it as a drug target because the function and behaviour of a protein depend completely on its structure. Since the 1950s, protein structure has been determined by experiment using X-ray crystallography (45, 46), and from the 1980s, NMR spectroscopy (47). However, it is time consuming, expensive and for some proteins, particularly those which do not easily crystallize, very difficult. With the advent of cheap DNA sequencing techniques the number of sequenced proteins ($\approx$ 5x10^7) (157) is significantly higher than the number of solved structures ($\approx 10^5$) (158, 159) and growing at a faster rate ($\approx 10^6$ new sequences in February 2014). Hence, an automated, *in silico* approach would be of enormous benefit for protein science.

Naïvely, one might not expect the problem to be too difficult; we have protein models and force fields, so we can initialize an unequilibrated protein and then evolve the state of system using molecular dynamics, enabling us to watch the protein fold into its native structure. However, there are problems with this technique. Firstly, the timescale over which proteins fold (typically μs to ms) is computationally infeasible for all-atom protein force fields, especially as the protein must be immersed in water for accurate results. For example, even with a special purpose-built computer, we can only successfully fold small proteins in atomic detail (51). Secondly, atomic protein force field energy functions are approximations which have been optimized for folded structures and it is unclear whether these approximations are justified for modelling unfolded proteins (130). This is particularly important as the free energy difference between the native structure and many non-native conformations is *relatively* small and could easily be eliminated when approximate energy functions are used. Due to these disadvantages more sophisticated techniques are usually employed.

The following sections describe the techniques that are currently used for protein structure prediction and how well they perform; related simpler prediction problems which aim to provide information about the structure, for example residue contact prediction; and finally a discussion of how correlated mutational analysis can be used in the context of protein structure prediction.

1.4.2 Existing Techniques

For prediction of protein structure, it is not just the physicochemical properties of the specific amino acid sequence of a protein which are relied upon; the repository of existing known protein structures (158) and the database of occurring primary sequences (157) can also be of tremendous importance. In fact, the most successful prediction techniques, 'template' methods, are a form of recognition and refinement rather than straightforward prediction.

Given a target sequence, a template method first seeks a homologous (similar) sequence for which a structure is already known. Dynamic programming sequence alignment programs, such as PSI-Blast (160), or Hidden Markov Models (161), are typically used to find this template structure. Once a template has been selected, it is used to build a model of the target protein which is then adapted, often using knowledge-based (162) or possibly physics-based refinements (163) in order to produce a final output.

Some of the best template methods include HHpred (164) and I-Tasser (162, 165), as shown by their performance in the bi-annual community-wide experiment, the Critical Assessment of Techniques for Protein Structure Prediction (CASP). Template-based modelling is maturing as a field and large proteins are now being predicted in all-atom detail relatively successfully (166). However, significant challenges still remain, both in choosing a template, especially when sequence similarity between target and template is low, and in successfully refining a correctly chosen template (166).

An alternative template-based approach is that of threading or fold recognition (167–170). The method does not rely on finding homologous sequences but instead tries to align the residues of the target sequence onto existing folded protein structures, comparing how well the residues 'fit' to the template, aiming to recognise the native fold. This method can work well as there seems to be a small number of protein folds found in nature, estimated to be of the order of a few thousand ((171) and references therein) and, as the CASP experiment has shown, the majority of newly determined proteins structures are found to have known folds.

In the case where either no template exists or one cannot be determined by sequence analysis, the prediction becomes significantly more difficult and currently only very small proteins can be determined to any degree of accuracy (172, 173). One of the most successful *ab initio* prediction algorithms is Rosetta, developed by David Baker's lab (174, 175). The algorithm starts with an extended chain, and small fragments, typically nine residues, of the chain are replaced by fragments from existing protein structures, chosen due to their residue similarities with the target sequence. The replacement is then accepted or rejected based on a Monte Carlo algorithm using a complicated knowledge-based energy function. Low energy conformations are then extracted, clustered and ranked, aiming to output the correct tertiary structure.

More novel approaches to the protein structure prediction have also been employed. For example, 'Fold it' is a game designed for the general public to try solve structure by manually optimizing a scoring function. It has had some success elucidating structures and is also used to learn how humans approach the prediction problem (176).

1.4.3 Protein Contact Prediction

Rather than determining the full tertiary structure of proteins directly, it is often easier to first predict more tractable information concerning the protein, such as its secondary structure. This information can then be used to guide the tertiary structure prediction.

Protein secondary structure prediction is a mature field; see (177) for a full overview. Typically, prediction involves taking the target sequence as an input to machine-learning algorithms, such as neural networks, first researched by Qian and Sejnowski in 1988 (178), and still popular today (179); support vector machines (180); or segmental Markov models (181, 182). Incorporating an aligned set of homologous sequences (a 'multiple sequence alignment') can vastly improve predictions (183). State of the art single sequence methods typically achieve 70% accuracy, whereas the addition of multiple sequence alignments can increase the accuracy up to nearly 80% (184). This is approaching the theoretical maximum of 88%; 100% cannot be achieved as different experimental models of the same protein can disagree with up to 12% of secondary structure assignments (177). Other protein features, such as residue solvent accessibility (see, for example (185)) and protein disulphide bonds (186) can also be predicted.

Protein Contacts

A particularly important subproblem is the prediction of protein contacts. A protein contact is a pair of residues sufficiently close to each other in the native state. For example, CASP deems two residues to be in contact if their C_β atoms are within 8Å of each other (187). Protein contacts are combined in a protein contact map, C_{ij}, a binary symmetric matrix where $C_{ij} = 1$ exactly when residues i and j are in contact (188). See Figure 1.11 for an example. Once again, machine learning algorithms, such as neural networks (189, 190), support vector machines (191) and random forests (192), have all been used to try to predict protein contacts. It is interesting to note that the latest CASP experiment was split into two separate categories: proteins for which a template exists (though challenging to identify) and proteins for which no template exists. The best contact prediction algorithm for the former category '*cheated*', in that it used the results of template-based prediction servers as inputs to its own contact prediction algorithm (187).

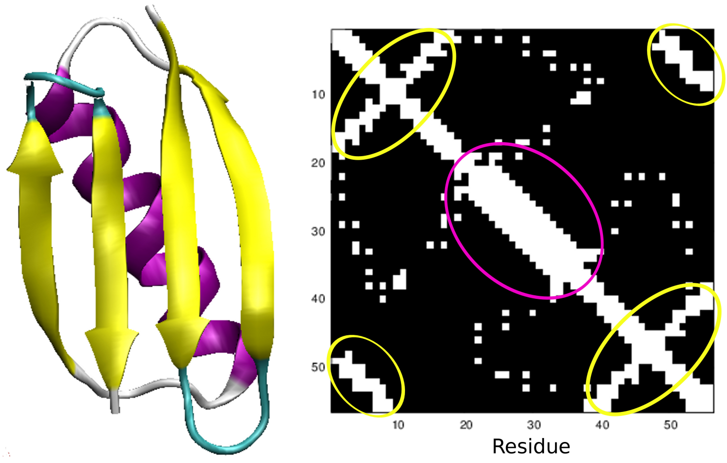


Figure 1.11: Left: The structure of protein G (1PGA). Right: the contact map of protein G. The white cells are residue contacts which occur when residue C_β (or C_α for glycine) atoms are within 8Å of each other. α-helices (highlighted in purple) and β-sheets (highlighted in yellow) can easily be read from the contact map and the remaining contacts describe how the secondary structure elements pack together to form this globular protein.

A discussion of another method to predict protein contacts, the use of correlated mutational analysis, is deferred until the following section. It is possible to use the predicted contacts to determine the full tertiary structure (193), but it has been shown, when using randomly chosen contacts, that between a quarter and two-fifths of native contacts are required (194–196).

Alongside the problem of predicting *general* protein contacts, a related problem is the prediction of protein β-contacts, or more generally the β-sheet topology of proteins. A β-contact is defined as a lateral pair of residues in interacting β-strands, as shown by Figure 1.12.

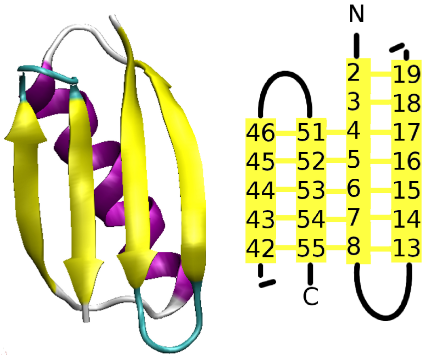


Figure 1.12: Left: The structure of protein G (1PGA). Right: The β-topology for protein G. The numbers are the positions in the sequence of the strand and the horizontal lines are β-contacts. For example residue 6 has two β-contacts: residues 15 and 53. Figure adapted from (197).

Unlike general protein contacts, β-contacts are subject to much more stringent constraints. For example, each residue can be involved in a maximum of two β-contacts, and due to the sequential nature of β-strands, given one or two β-contacts, many more are then forbidden. This extra structure can be incorporated into inference algorithms as prior knowledge. The prediction of β-contacts can be used to directly aid protein structure prediction (116, 198), in designing proteins (199, 200) and understanding protein folding pathways (201, 202).

A wide variety of machine-learning algorithms have been developed for β-contact prediction. BetaPro, a 2D recurrent neural network, the work of Cheng and Baldi (203), is noteworthy for being the first algorithm which takes the global nature of β-topologies into account. A variety of other approaches are also taken, for example, the use of statistical potentials (204), information theoretic approaches (205), hybrid neural network-probabilistic models (206), segmental Markov models (182) and Markov logic networks (207).

1.4.4 Correlated Mutational Analysis

Given a primary sequence of a protein, one can generate an aligned set of homologous sequences (a ‘multiple sequence alignment’, MSA), first by using PSI-Blast (160) to generate a list of proteins with similar sequences and then using an alignment program, see for example (208, 209), to generate the MSA. Figure 1.13 shows a toy MSA which is referred to throughout this section.

The MSA of a protein contains information about proteins evolutionarily related (and hence probably structurally related) to it. Therefore the MSA is usually an integral input into the various machine learning algorithms described above, and using it significantly improves their performance. As an illustrative example, given only the target sequence in Figure 1.13, it is unclear whether there is a disulphide bridge, and if so, whether it is between cysteines in positions (1,5), (1,10) or (5,10). In many cases, if there is a disulphide bridge, it is essential for the structural stability of the protein, and therefore is usually highly conserved between homologous proteins. In our example, using the MSA, it is likely there is a disulphide bridge between residues 1 and 5, as only these cysteines are highly conserved.

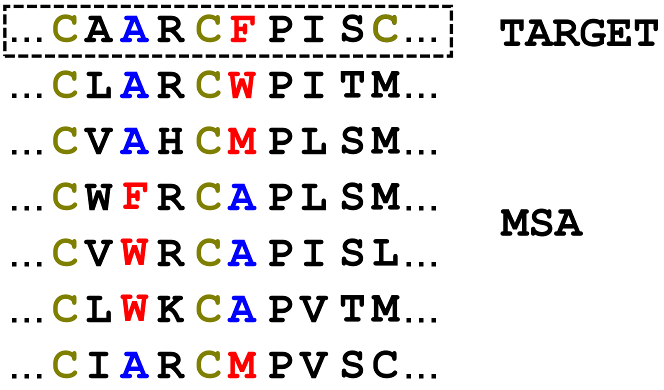


Figure 1.13: The top line is part of the primary sequence of a (toy) protein and the following lines show its MSA. Each row represents a different protein and each column a residue position. The conserved cysteines (C) in columns 1 and 5 strongly suggest a disulphide bridge between these residues; see the text for more details. It is also likely that residues 3 and 6 have co-evolved, as whenever alanine (A) in column 3 mutates into a larger residue, the large residue in column 6 compensates by mutating into the smaller residue, alanine; see the text for more details. We define $f_i(A_i)$ as the observed frequency of residue A_i in position i and $f_{ij}(A_i, A_j)$ as the observed frequency of both A_i in position i and A_j in position j. In this case $f_3(A) = 4/7$, $f_7(P) = 1$ and $f_{3,8}(A, I) = 2/7$.

In contrast to other methods, which typically use the MSA as an input into a machine learning algorithm, the technique of correlated mutational analysis (CMA) looks at correlations between the columns of an MSA in an attempt to predict residues that have co-evolved. Co-evolution may imply structural proximity in the folded state, for example, if a residue mutates into a larger one, a compensatory mutation, reducing the size of a structurally proximal residue may be required to maintain the viability of the fold; hence these two residues have 'co-evolved'. An example is shown by columns 3 and 6 in Figure 1.13.

Although these techniques date back over 20 years (210, 211), with the increasing number of sequences now available, CMA has recently received a lot of attention (212, 213). Early CMA methods were developed using Pearson correlation coefficients (210) and perturbation methods (214), but we focus on methods which use mutual information (MI) to detect correlated pairs of residues. Given an MSA, we define $f_i(A_i)$ as the observed frequency of residue A_i in position i and $f_{ij}(A_i, A_j)$ as the observed frequency of both A_i in position i and A_j in position j; see Figure 1.13 for examples. The MI of residues i and j, $M(i, j)$ is defined as

$$M(i,j) = \sum_{A_i} \sum_{A_j} f_{ij}(A_i, A_j) \ln \left(\frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right)$$

and a large $M(i, j)$ implies residues i and j are correlated, which *may* imply a protein contact.

Unfortunately MI is often a poor predictor of residue contacts for a number of reasons. For example, background variability of the columns can skew the results; typically columns of the MSA with higher variability have higher MI scores. Also, if a large number of near identical sequences are included in the MSA, for example if the protein evolves slowly in one family of organisms, then the signal from this family will be amplified. There has been work adapting MI to compensate for these problems (215–217). However, the pairwise nature of MI means it is unable to distinguish between transitive correlations and true protein contacts: if residue pairs (i,j) and (j,k) are in contact and their MSA columns are correlated, residue pair (i,k), which is not a contact, is also likely to be correlated and this cannot be distinguished by pairwise statistics such as MI.

In an attempt to disentangle direct and indirect correlations, in under-recognised work, Lapedes and coworkers developed a statistical model of the entire MSA (218). Recently this model has been revived and made computationally less expensive and has successfully predicted protein-protein interaction contacts (219) and protein residue contacts to sufficient accuracy to determine tertiary structure of both globular (193) and even membrane proteins (220).

Unlike pairwise methods, in this case the entire MSA is modelled by assigning probability mass over all (fixed length) sequences, even those which have not been observed. This probability distribution, $\mathfrak{P}(\mathbf{A})$, matches the low order empirical moments of the observed MSA. Specifically $P_i(A_i) = f_i(A_i)$ and $P_{ij}(A_i, A_j) = f_{ij}(A_i, A_j)$ where $P_i(.)$ is the marginal distribution for position i and $P_{ij}(.,.)$ is the joint marginal distribution for positions i and j. $\mathfrak{P}(\mathbf{A})$ is then fully defined following the principle of maximum entropy (221). Specifically $\mathfrak{P}(\mathbf{A})$ is the distribution with maximizes entropy, $S(\mathfrak{P}) = -\sum_{\mathbf{A}} \mathfrak{P}(\mathbf{A}) \log \mathfrak{P}(\mathbf{A})$, whilst satisfying the above constraint.

$\mathfrak{P}(\mathbf{A})$ can be expensive to compute; message passing (219), perturbation methods (222) and mean-field approximations (193) have all previously been used, yet it is interesting to note that this computation happens to be ideally suited to the contrastive divergence method mentioned earlier. However, once $\mathfrak{P}(\mathbf{A})$ is known, the direct information (DI), a statistic analogous to MI, can be computed. DI is an attempt to ascertain the strength of direct correlations between pairs of residues taking into account the global nature of the MSA, through the use of $\mathfrak{P}(\mathbf{A})$. Further details and discussions can be found in Chapter 4.

Unfortunately, as the number of sequences in a MSA decreases, CMA methods lose their effectiveness (211), and therefore authors have typically focussed on a small number of proteins for which large high quality MSAs exist. In these cases a structure representative of the family almost always exists and template-based methods could successfully have been used.

An interesting question is, when the MSA is small, as is often the case when no template exists, does it still contain co-evolution information and can CMA methods still be used to extract it? A promising idea is to combine prior knowledge about protein contacts into the CMA method; for example Burger and van Nimwegen have developed a dynamic Bayesian network to predict protein contacts which is dramatically improved when used with an informative prior which takes into account, for example, the fact that contact probability strongly depends on sequence distance (223).

1.5 Thesis Overview

This is a thesis principally based on published work. Chapters 2–4 contain published research presented as journal articles and Chapter 5 includes unpublished work written in paper format. Each chapter is prefaced by a summary and context of the research work and a statement describing the contribution of the candidate to the work.

Chapter 2 describes the first application of the nested sampling algorithm to a biophysical system. The nested sampling algorithm is parallelized and applied to the CRANKITE coarse-grained protein model, which has also been improved with the addition of side chain γ-atoms and a hydrophobic energy function. The potential energy landscapes of three small proteins are explored and energy landscape charts are generated giving a large-scale visualization of the potential energy surface. Finally, the nested sampling algorithm is compared to parallel tempering, with both methods calculating the heat capacity of a polyalanine α-helix.

CRANKITE is further improved in Chapter 3, where the contrastive divergence algorithm is used to learn the parameters for the Lennard Jones (LJ) form of the van der Waals potential. Unlike previous maximum likelihood optimized models, the CRANKITE force field is shown to be transferable to proteins not in the training set. The more expensive and accurate LJ potential is shown to be more effective in modelling the dense packing of the hydrophobic core compared to a more simple steric repulsion function. The importance of correctly optimizing parameters, rather than taking literature values or those found in other force fields, is also discussed.

CRANKITE can also been used when the native structure of a protein is unknown in order to predict protein structure. However, β-contacts are required as an input for the model. For unknown structures these contacts must be predicted, and in Chapter 4 a new β-contact prediction algorithm is described which incorporates a powerful correlated mutation statistic, the direct information (DI). Unlike in previous work, CD is used to estimate the maximum entropy distribution from which the DI is derived. Furthermore, unlike the majority of correlated mutation analysis research, proteins with large high-quality multiple sequence alignments are not 'cherry-picked' for analysis, but instead a standard dataset of 916 proteins used to benchmark β-contact prediction algorithms is used. The DI, which is often noisy for proteins with small MSAs, is coupled to an empirical Bayes β-sheet model. The strong constraints and prior knowledge associated with β-contacts are ideally suited for prediction using a method that incorporates the often noisy DI. The developed method is competitive with existing state of the art prediction methods whilst requiring fewer inputs.

Standard nested sampling is a Monte Carlo sampling algorithm, and for coarse-grained models such as CRANKITE, MC moves such as crankshaft rotations can be used successfully. However, for more complicated all-atom models, which have more degrees of freedom per residue, additional moves such as angle bending and side chain rotations must be included. These MC moves, particularly at low temperatures, are less efficient, and so molecular dynamics simulations are normally used for all-atom models. Therefore, in order for the nested sampling algorithm to gain popularity within the computational structural biology community, it is necessary to adapt the nested sampling algorithm to incorporate MD, whilst still retaining its inherent advantages.

Recently, Skilling has developed *Galilean* nested sampling (224), a version of nested sampling which can be implemented within a MD framework. In Chapter 5 Galilean NS has been implemented for use with the Amber MD software package. The algorithm is used to sample an all-atom model of alanine dipeptide, producing estimates of free energies and heat capacities. The performance of the new algorithm is compared to that of the standard REMD approach.

Finally Chapter 6 concludes the work, summarizing the thesis and describing possible avenues of further work.

Chapter 2

Nested Sampling for a Coarse-Grained Protein Model

Nested sampling is a Bayesian sampling algorithm which is particularly efficient at sampling systems which undergo a first order phase transition. The algorithm has previous been applied to calculate the heat capacities of small Lennard Jones (LJ) clusters, where the algorithm was shown to be an order of magnitude more efficient than the standard parallel tempering approach.

In this chapter we parallelize the nested sampling algorithm and, for the first time, apply it to a biophysical system. We use the algorithm to explore the potential energy surface of three small globular proteins using a coarse-grained protein model, CRANKITE. We have improved the existing CRANKITE model by more accurately modelling amino acid side chains, including a specific hydrophobicity energy term and modifying the functional form of existing energy terms.

In order to test algorithm efficiency, we calculate the heat capacity of a 16-residue polyalanine helix using both parallel tempering and nested sampling. Unlike phase transitions of LJ clusters, the transition from coil to helix occurs over a large temperature range and so parallel tempering is able to successfully sample the system. Therefore, we find nested sampling and parallel tempering to be of similar efficiency for this system. However, it has been shown that larger proteins undergo a first order phase transition and in these cases, we would expect nested sampling to outperform parallel tempering.

The nested sampling algorithm provides conformations taken from throughout the entire potential energy surface, and we cluster these samples and use them to generate energy landscape charts. These give a large-scale visualization of the potential energy surface and provide insights into the force field and model used.

For example, using CRANKITE to study protein G, these charts show a high energy barrier between conformations with the helix on the same side of the β-sheet as in the native structure and the helix on the other side of the sheet. The top-down approach of nested sampling ensures that both basins are explored. When using parallel tempering, it is necessary for the highest temperature replica to be able to cross all energy barriers and it can be challenging to ensure this is the case because the height of energy barriers is not known *a priori.*

Although nested sampling is *athermal*, the samples can be weighted to generate a distribution of conformations at any temperature. We take the room-temperature distribution of conformations for protein G and compare the distribution to the samples found when using rigidity analysis.

Rigidity analysis uses the native structure of a protein to build an elastic network model of the protein, which is used to determine the rigid and flexible regions of the protein. Normal mode projection is then used to move the flexible regions of the protein in order to estimate the conformational changes which are accessible at a given temperature. We find good agreement between these two, very different, coarse-grained models.

We have shown that nested sampling is a promising algorithm for Monte Carlo sampling of coarse-grained protein models.

2.1 Contribution

The initial idea was formulated by Wild. A serial implementation of the nested sampling algorithm, using an earlier version of CRANKITE, together with the generation of 'posterior' energy landscape charts, was submitted by Burkoff to the University of Warwick towards an MSc in Scientific Computing.

Wild provided general supervision throughout the project. Burkoff parallelized and implemented the parallel nested sampling algorithm. Várnai, with the help of Burkoff, developed the updated CRANKITE model and force field. Burkoff implemented the 'prior' energy landscape charts under the supervision of Várnai. Burkoff calculated the heat capacity curves and benchmarked the results against parallel tempering. Wells implemented the rigidity analysis of protein G and wrote Section 3 of the Supporting Material.

The initial draft of the paper was written by Burkoff, with Wild and Várnai contributing to the revision of the paper before submission. Burkoff also drafted a response to referees with input from Wild and Várnai during the peer review process and revised the paper before publication.

The article was originally published by Cell Press: Burkoff, N. S., C. Várnai, S. A. Wells and D. L. Wild. Exploring the Energy Landscapes of Protein Folding Simulations with Bayesian Computation. *Biophysical J.* (2012) 102 878–886.

Exploring the Energy Landscapes of Protein Folding Simulations with Bayesian Computation

Nikolas S. Burkoff,[†] Csilla Várnai,[†] Stephen A. Wells,[‡] and David L. Wild[†*]
[†]Systems Biology Centre and [‡]Department of Physics and Centre for Scientific Computing, University of Warwick, Coventry, United Kingdom

ABSTRACT Nested sampling is a Bayesian sampling technique developed to explore probability distributions localized in an exponentially small area of the parameter space. The algorithm provides both posterior samples and an estimate of the evidence (marginal likelihood) of the model. The nested sampling algorithm also provides an efficient way to calculate free energies and the expectation value of thermodynamic observables at any temperature, through a simple post processing of the output. Previous applications of the algorithm have yielded large efficiency gains over other sampling techniques, including parallel tempering. In this article, we describe a parallel implementation of the nested sampling algorithm and its application to the problem of protein folding in a Gō-like force field of empirical potentials that were designed to stabilize secondary structure elements in room-temperature simulations. We demonstrate the method by conducting folding simulations on a number of small proteins that are commonly used for testing protein-folding procedures. A topological analysis of the posterior samples is performed to produce energy landscape charts, which give a high-level description of the potential energy surface for the protein folding simulations. These charts provide qualitative insights into both the folding process and the nature of the model and force field used.

INTRODUCTION

Approximately 50 years ago, Anfinsen and colleagues (1) demonstrated that protein molecules can fold into their three-dimensional native state reversibly, leading to the view that these structures represented the global minimum of a rugged funnel like energy landscape (1–3).

According to the hierarchical folding theory of Baldwin and Rose (4,5), a protein folds by first forming local structural elements, namely, α-helices and β-strands. These secondary structure elements then interact with each other, resulting in the formation of the folded protein. The formation of local structural elements reduces the entropy of the protein (for example, the side chains of helical residues are strongly constrained by the rest of the helix). This loss of entropy is compensated by favorable short-range interactions, including hydrogen bonding and desolvation of backbone polar groups. This is considered to be a fundamental property of proteins, and any model system attempting to simulate protein folding should mimic this property.

Although there has been recent evidence of hierarchical folding in long timescale molecular dynamics simulations made possible by the use of custom designed supercomputers (6), simplified Gō-type models remain an important class of protein models in the investigation of energy landscapes. Gō models assume that nonnative interactions do not contribute to the overall shape of the folding energy surface (7,8). In this work we use an extended Gō-type model, in which a Gō potential captures interactions between contacts of the native state of the protein, but attractive nonnative interactions are also permitted (for example, hydrogen bonds can form between residues that are not in contact in the native state). This addition allows us to explore a more realistic rugged energy landscape compared to the "perfect funnel" found in a standard Gō model (8), while maintaining the ability to perform simulations with limited computational resources.

The energy landscapes of protein-folding simulations are most commonly visualized in terms of two- or three-dimensional plots of microscopic or free energy versus a reaction coordinate, such as the fraction of residue contacts in common with the native state or the root mean-square deviation (RMSD) between a given conformation and the native state (9,10). Originally developed for reduced lattice models, these approaches have since been used for all-atom off-lattice simulations, although, in these more realistic models, they offer only an indirect visualization of the energy landscape at a single scale (11). Projection into the space defined by principal components analysis of the contact map has also been used to provide a two-dimensional visualization of the energy surface (12). Techniques adapted from robotic motion planning have been used to provide a probabilistic roadmap of protein folding, which may be mapped onto a conceptual drawing of the potential energy surface (13). Protein potential energy surfaces and folding funnels have also been visualized by disconnectivity graphs (14) and scaled disconnectivity graphs (15,16). Although these latter methods have the advantage of providing a visualization of the whole energy landscape, they rely on creating a large database of local energy

Submitted September 13, 2011, and accepted for publication December 27, 2011.

*Correspondence: d.l.wild@warwick.ac.uk

Editor: Michael Levitt.

0006-3495/12/02/0878/9 $2.00

doi: 10.1016/j.bpj.2011.12.053

minima of the surface, and are therefore impractical for large systems; they also do not provide information about the entropy of the system (which governs the widths of the conceptual protein-folding funnel).

The funnel like nature of the energy landscape provides a challenging conformational space for computer simulations to explore, because only an exponentially small number of conformations have low energy and low entropy and are found toward the bottom of the funnel; the system also undergoes a first-order phase transition as the protein collapses into its native state. In this work, we use nested sampling to explore the energy landscapes of protein folding simulations. Nested sampling is a Bayesian sampling technique introduced by Skilling (17,18), designed to explore probability distributions where the posterior mass is localized in an exponentially small area of the parameter space. It both provides an estimate of the evidence (also known as the marginal likelihood, or partition function) and produces samples of the posterior distribution. Nested sampling offers distinct advantages over methods such as simulated annealing (19), Wang-Landau sampling (20), parallel tempering (replica exchange) (21), and annealed importance sampling (22), in systems characterized by first-order phase transitions (17,23). The technique reduces multidimensional problems to one dimension and has a single key parameter in the trade-off between cost and accuracy. The calculation of free energies by thermodynamic integration (24) and thermodynamic observables, such as heat capacities, typically involves multiple simulations at different temperatures. Nested sampling provides an efficient framework for computing the partition function and hence thermodynamic observables at any temperature, without the need to generate new samples at each temperature. Hence, it allows us to directly investigate the macroscopic states of the protein-folding pathway and evaluate the associated free energies. Nested sampling has previously been used in the field of astrophysics (25) and for exploring potential energy hypersurfaces of Lennard-Jones atomic clusters (23), yielding large efficiency gains over parallel tempering. Its use in this article represents, to our knowledge, the first application of this technique to a biophysical problem.

MATERIALS AND METHODS

In general, the energy of a polypeptide, $E(\Omega,\theta)$, is defined by its conformation, Ω, and arbitrary interaction parameters, θ. These interaction parameters may be as diverse as force constants, distance cutoffs, dielectric permittivity, atomic partial charges, etc. This energy, in turn, defines the probability of a particular conformation, Ω, at inverse thermodynamic temperature β via the Boltzmann distribution

$$P(\Omega,\theta|\beta) = \frac{1}{Z(\theta,\beta)}\exp[-E(\Omega,\theta)\beta], \quad (1)$$

$$Z(\theta,\beta) = \int d\Omega \exp[-E(\Omega,\theta)\beta], \quad (2)$$

where $Z(\theta, \beta)$ is the partition function (or evidence, in Bayesian terminology). In the following, energy is expressed in units of RT, the product of the molar gas constant and absolute temperature and $\beta = 1/RT$.

In Bayesian statistics, with θ an unknown parameter, D the observed data, and H the underlying model or hypothesis, we have the following relation (Bayes' rule)—posterior × evidence = likelihood × prior—

$$\mathcal{P}(\theta|D,H)Z = \mathcal{P}(D|H,\theta)\mathcal{P}(\theta|H),$$

where Z, the evidence, is defined as

$$Z = \int \mathcal{P}(D|H,\theta)\mathcal{P}(\theta|H)d\theta.$$

Nested sampling provides an algorithm for estimating the evidence, $Z = \mathrm{P}(D|H)$, and the procedure additionally explores the posterior distribution, allowing its properties to be estimated.

Procedure

We define $X(\lambda) = \lambda$ to be the proportion of the prior distribution with likelihood $L(X) > \lambda$. Then, following Skilling (17), the evidence is

$$Z = \int_0^1 L(X)dX,$$

where $L(X(\lambda)) = \lambda$ and $dX = \pi(\theta)d\theta$, with $\pi(\theta)$ the prior distribution. Fig. 1 shows the graph of L against X (this is not to scale, as normally the bulk of the posterior is in an exponentially small area of the phase space). L is a decreasing function of X, as the restriction on the likelihood becomes tighter as λ increases. The area under the curve is Z. The nested sampling procedure estimates points on this curve (see Algorithm, below) and then uses numerical integration to calculate Z.

Algorithm

1. Sample (uniformly, with respect to the prior distribution) K points of the parameter space $\{\theta_1 ... \theta_K\}$, i.e., the "active list"; then calculate their likelihoods: $\{L(\theta_1),...,L(\theta_K)\}$.
2. Take the sample point with the smallest likelihood; save it as (L_1, X_1) (see below for an estimate of X); remove this point from the active list.

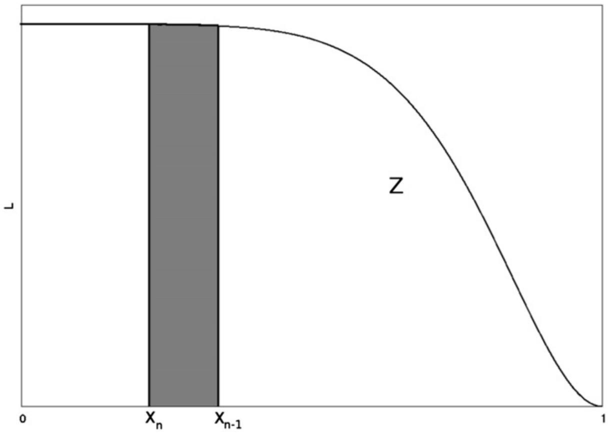


FIGURE 1 Evidence Z is the area under the function $L(X)$. The sample θ_n represents $X_{n-1} - X_n$ of the phase space volume; the proportion of the x axis is shaded. Its weighting for the posterior is L_n $(X_{n-1} - X_n)/Z$; the proportion of Z is shaded.

3. Generate a new point θ sampled uniformly (with respect to the prior distribution) from those points with likelihood $L(\theta) > L^* = L_1$; then add it to the active list.
4. Repeat Steps 2 and 3, generating (L_2, X_2), (L_3, X_3),…,(L_1, X_i),… .

X_1 is located at the largest of N numbers uniformly distributed on $(0, X_0)$, where $X_0 = 1$. Skilling (17) suggests using the expected value of the shrinkage ratio, X_i/X_{i-1}, to estimate X_i (the estimate of X for iteration i), where X_i is the largest of N numbers uniformly distributed on $(0, X_{i-1})$. The shrinkage ratio has the probability density function $f(t) = Kt^{K-1}$, with mean and standard deviation $\log(t) = (-1 \pm 1)/K$, and, as each shrinkage ratio is independent, we find, if uncertainties are ignored,

$$\log(X_i) = \left(-i \pm \sqrt{i}\right)/K \Rightarrow X_i \approx \exp(-i/K).$$

It is also possible to use the arithmetic expected value to estimate X_i (26). This implies that $X_i = \alpha^i$, where $\alpha = K/(K+1)$. In the limit of large K, these two approaches are identical and henceforth we will use $\alpha = \exp(-1/K)$ or $K/(K + 1)$, and $X_n = \alpha^n$.

Parallel nested sampling

For high-dimensional systems, sampling uniformly (conditional upon the likelihood being above a fixed value, L^*) is not computationally tractable. In this case, a Markov chain can be used to explore the parameter space (22). To generate a new point, one of the active set of points (not necessarily the one with the lowest likelihood) is chosen to be the start of a short Monte Carlo (MC) run, with all moves that keep the likelihood above L^* being accepted.

Starting the MC run from a copy of one of the points of the active set, chosen at random, is crucial to nested sampling. Suppose we have a bimodal likelihood function. Once L^* is sufficiently high, the region of the parameter space the chain is allowed to explore will no longer be connected; it will have two disconnected components. Without copying, all active points that enter the subordinate component will be trapped there. With copying, provided at least one enters the dominant mode, then as L^* increases, active points in the subordinate mode will be replaced by ones from the dominant mode. This is particularly important for likelihood functions for which the dominant mode splits again at a higher likelihood. In general, for a given K, if the relative phase space volume of a mode is <1 / K in comparison to the rest of the space at the splitting likelihood, the chances of nested sampling exploring the mode is small (23). Therefore, the parameter K controls the resolution of the exploration.

The number of trial MC moves per nested sampling iteration, m, is another key parameter when using nested sampling for higher dimensional systems. If m is too small, the parameter space is inadequately explored; new active set samples and the current conformations they are copied from remain very similar. Setting m too high results in longer than necessary runtimes, as conformations partway through the MC run are already sufficiently different from their starting positions. Hence, K controls which regions of the parameter space are available to the algorithm and m controls how well these regions are explored.

We parallelized the nested sampling algorithm by removing the P points with the lowest likelihood at each nested sampling iteration, one for each processor used. Each processor then runs its own independent MC simulation to replace one of the removed points. For post processing, at each iteration we only store the point that has the P^{th} lowest likelihood and adjust α accordingly; $\alpha = 1 - P/(K + 1)$.

Running a parallel nested sampling algorithm with K points explores the parameter space more effectively than P serial nested sampling simulations each with K/P points in the active set, while requiring equal computational resources. Consider a likelihood function, which splits n times in the dominant mode (i.e., contains the majority of the evidence), with the probabilities of an exploratory active point falling into the dominant mode being $W_1, W_2, \ldots W_n$ at the critical likelihood (which is the likelihood of splitting). Defining success as exploring the dominant mode at the n^{th} split in at least one simulation, it can be shown, using an argument similar to that of Sivia and Skilling (18), that

$$\mathbb{P}(\text{success}|\text{one simulation with } K \text{ points}) = \Pi_g\left[1-\left(1-W_g\right)^K\right] \tag{3}$$

and

$$\mathbb{P}(\text{success}|P \text{ simulations with } K/P \text{ points}) = 1 - \left(1 - \Pi_g\left(1 - \left(1 - W_g\right)^{K/P}\right)\right)^P.$$

For example, if $n = 2$, $W_1 = W_2 = 0.1$, $K = 32$, and $P = 4$, then $\mathbb{P}(\text{success}|\text{parallel}) = 0.933$ and $\mathbb{P}(\text{success}|\text{serial}) = 0.792$.

Posterior samples

The sample points removed from the active set, labeled θ_1, θ_2,…, say, can be used to estimate properties of the posterior distribution. Sample point θ_n represents

$$\omega_n = X_{n-1} - X_n$$

of the phase space volume (with respect to the prior distribution) and hence

$$\chi_n = \frac{(X_{n-1} - X_n)L(\theta_n)}{Z}$$

is the relative volume of the posterior space that θ_n represents; see Fig. 1.

In the case of a Boltzmann distribution, at inverse temperature β, $L(\theta_n) = \exp(-E_n\beta)$ and hence, by calculating $\chi_n(\beta)$, a single nested sampling simulation can provide the expectation value of any thermodynamic observable, such as heat capacity, at any temperature. Given a property $Q(\theta|\beta)$ of the posterior,

$$\mathbb{E}(Q|\beta) \approx \sum_i \chi_i(\beta) Q(\theta_i). \tag{4}$$

In energetic terms, the nested sampling scheme is built from a set of decreasing energy levels, $\{E_n\}$, with the energy of conformation Ω_n given by Eq. 5. Each energy level has an associated weight, which is also decreasing. At each energy level, a set of K sample points (or conformations), $\{\Omega^i{}_n\}$, is obtained by uniform sampling from the energy landscape below E_n: $\Omega^i{}_n \sim U(\Omega\text{: } E(\Omega) < E_n)$. After every iteration, a new lowest energy level E_{n+1} is defined to be at a fixed fraction, α, of the current energy distribution. In this way, a fraction α^n of the whole phase space has energy below E_n, and a fraction α^{n+1} has energy below E_{n+1}. The phase space volume will therefore shrink exponentially, by a factor of α, with every nested sampling iteration, and the algorithm is able to locate exponentially small regions of phase space.

The protein model

The polypeptide model we use is adapted from our previous published work (27–30). It is fully described in the Supporting Material and a summary is provided below.

Our polypeptide model features all-atom representations of the polypeptide backbone and β-carbon atoms. Other side-chain atoms are represented by one or, in the case of branched side chains, two pseudoatoms, following Srinivasan and Rose (31).

For a given protein sequence, R, the Boltzmann distribution defines the probability, $P(R, \Omega|\beta)$, of it adopting a particular conformation, Ω, at inverse thermodynamic temperature β. This probability can be factorized into the product of the sequence-dependent likelihood for a given conformation and the prior distribution of conformations, $P(R, \Omega) = P(R|\Omega)P(\Omega)$. This can be rewritten in energetic terms as

$$E(R, \Omega) = -\ln P(R|\Omega) + E(\Omega), \tag{5}$$

where sequence-dependent and sequence-independent contributions to the energy are separated. We assume that the sequence-independent term, $E(\Omega)$, is defined by short-range interactions among the polypeptide backbone, β-carbon, and pseudo-atoms. At room temperature, van der Waals repulsions and covalent bonding between atoms are extremely rigid interactions that contribute to this energy. Another large contribution comes from hydrogen bonding, but the magnitude of this interaction is vaguely understood. The sequence-dependent part of the potential (the negative log-likelihood) can be approximated by the pairwise interactions between side chains, which make the largest contribution to this term. In this work, these interactions are modeled by a Gō-type potential based on a regularized native contact map (27), which contains lateral contacts in parallel and antiparallel β-sheets and contacts between residues i and $i + 3$ in α-helices (32,33). Our model also includes a hydrophobic packing term; hydrophobic side chains coming into contact with hydrophobic or amphipathic side chains are rewarded with a decrease in energy (31). The force constants for these side-chain interactions, as well as backbone hydrogen bonding, are optimized using a novel statistical machine learning technique (29).

Nested sampling is initialized with K conformations, uniformly distributed over the space of dihedral angles (i.e., every $\phi_i, \psi_i \sim U[-180°, 180°]$). To generate new sample points we use our implementation of an efficient Metropolis Monte Carlo (MMC) algorithm (28,30), which relies on local Metropolis moves, as suggested in earlier studies (34). In contrast to other programs that rely on local Metropolis moves in the space of dihedral angles, our sampler utilizes local crankshaft rotations of rigid peptide bonds in Cartesian space. An important feature of our model is the elasticity of the α-carbon valence geometry. With flexible α-carbon valence angles, it becomes possible to use crankshaft moves inspired by earlier MMC studies of large-scale DNA properties. The amplitudes of proposed crankshaft rotations were chosen uniformly from $[-\alpha_0, \alpha_0]$ where, at every 2000 nested sampling iterations, α_0 (the maximum allowed proposed amplitude) was recalculated, attempting to keep the acceptance rate at 50% (the trial MC moves used for this calculation were then ignored).

We ran simulations until $Z(\beta)$ converges for $\beta = 1$ ($T = 25°$C), which implies that we have sampled from the thermodynamically accessible states for all temperatures smaller than β ($>T$). The nested sampling algorithm marches left across the x axis of Fig. 1. The step size is constant in log X and the larger the K, the smaller the step size. For a given protein and β, we find that simulations terminate at approximately the same point on the x axis (for protein G, with $\beta = 1$, this is $\sim e^{-440}$). This implies that the total number of iterations is proportional to K, and the total number of MC moves is proportional to mK. The results for protein G shown below are from a simulation with $K = 20{,}000$ and $m = 15{,}000$, which used 32 processors (Xeon X5650; Intel, Santa Clara, CA), had 1.38×10^{11} MC moves, and took ~22 h.

Energy landscape charts

We use the algorithm recently introduced by Pártay et al. (23), which uses the output of a nested sampling simulation to generate an energy landscape chart, facilitating a qualitative understanding of potential energy surfaces. It has the advantage of showing the large-scale features of the potential energy surface without requiring a large number of samples.

The output of a nested sampling simulation is a sequence of sample points with decreasing energy. Each sample point (conformation), Ω_n, represents $\omega_n = \alpha^{n-1} - \alpha^n$ of the phase space and has energy $E_n(\Omega_n)$. A metric defining the distance between two conformations is required, and using this, a topological analysis of the sample points is performed. As the metric, we use the root mean-square deviation of the backbone and side-chain nonhydrogen atoms of a pair of conformations; that is, the sum of the Euclidean distances of corresponding atoms after the two conformations have been translated and rotated in space to minimize the overall distance.

A graph $\mathcal{G}$ is created with the sample points as nodes and arcs joining a sample to the k nearest samples that have higher energy. In this work, k is chosen to be 15 throughout. We then start with an empty graph ($\mathcal{G}'$), adding nodes one at a time (starting with the lowest energy) to gradually rebuild $\mathcal{G}$.

Energy landscape charts are produced with energy on the vertical axis, and, at a given energy E_n, the width of the chart is proportional to the sum of the weights of the points below that energy (i.e., $\omega_n + \omega_{n+1} + \ldots$), that is, the available phase space volume in the prior space, contained at $<E_{n.}$ On the horizontal scale, the chart is split into different basins corresponding to the disconnected subgraphs that exist when sample n is added to $\mathcal{G}'$. The relative widths of the basins is given by the ratio of the sum of the weights of the sample points in the disconnected subgraphs. The ordering of the basins horizontally is arbitrary. Due to the rapid shrinking of the available phase space volume with decreasing energy, for better visualization, a horizontal scaling is applied by an exponential function of the energy, similar to Pártay et al. (23). The energy landscape chart represents a potential energy landscape for the system.

We also use a variant of the energy landscape charts where the width of the chart is proportional to the sum of the posterior weights, $\chi_n = \omega_n \exp(-E_n\beta)/Z(\beta)$, i.e., ($\chi_n + \chi_{n+1} + \ldots$), at inverse temperature β. Hence, the relative widths of the basins correspond to the probabilities of adopting a conformation from one basin or another. These energy landscape charts, therefore, represent the energy landscape as it is experienced by the protein at inverse temperature β. In the following, the two versions will be referred to as prior and posterior energy landscape charts, according to the weights used in the calculation of their basin widths.

RESULTS

To validate the nested sampling procedure, we simulated the folding of an isolated 16-residue polyalanine β-hairpin. We then conducted folding simulations on a number of small proteins that are commonly used for testing protein folding procedures: protein G (PDB code 1PGA), the SH3 domain of Src tyrosine kinase (PDB code 1SRL), and chymotrypsin inhibitor 2 (PDB code 2CI2).

Isolated polyalanine β-hairpin

We used a Gō-like potential to simulate the folding of an isolated 16 residue polyalanine β-hairpin. Fig. S1 in the Supporting Material (*bottom panel*) shows a snapshot of five (equally spaced along the log(X) axis) conformations from a single simulation with $K = 1000$, $m = 2500$ (a total of 1.12×10^8 MC moves). At the beginning there is a rapid decrease in energy, moving from extended conformations (at first those with van der Waals collisions) to hairpinlike structures (*A–C*). The final part of the simulation moves through the exponentially small volume of the phase space containing hairpinlike structures, gradually decreasing in energy toward a fully formed hairpin (*D* and *E*).

We used the hairpin to check the behavior of the nested sampling procedure: Fig. S1 (*top panel*) shows how α_0 (the maximum proposed crankshaft rotation amplitude) varies with the energy threshold for a simulation with

$K = 1000$. As lower energy is reached, α_0 is reduced to keep the acceptance rate near 0.5. Fig. S1 (*second panel*) shows the acceptance rate. Fig. S1 (*third panel*) shows the difference between the start and end points of a single MC chain, specifically the drift per dihedral angle, where the drift is the L_2-norm of the dihedral angles.

The protein model used stabilizes room temperature secondary structure formation; it folds isolated helices and hairpins very effectively. This is reflected in the energy landscape charts that consist of a single funnel (not shown).

Fig. S2 (*top*) shows the time evolution of the dihedral angles of four residues of the 16-residue polyalanine. The formation of the hairpin can be clearly seen. For example, the dihedral angles of the residues in the strands 4 and 11 converge to the standard β-sheet area of the Ramachandran plot. The Gō-like potential used was designed for a hairpin with a two-residue turn, and this is found to be the case. The dihedral angles of the turn residues 8 (60 ± 15, -90 ± 30) and 9 (-150 ± 30, 0 ± 30) are closest to the values of a type II$'$ turn ((60, -120) and (-80, 0)) (35). Fig. S2 (*bottom*) shows the energy of the snapshots (*right-hand axis*) for nested sampling plotted against time. The monotonic decrease of the energy over a very large energy range allows us to view the formation of the hairpin.

Due to the nature of the model used, the folding pathway of the hairpin is relatively simple to sample, and parallel tempering can also successfully fold the hairpin. However, in this case, we need a very large temperature range to explore the whole parameter space and view the folding pathway in its entirety. For example, Fig. S2 (*bottom*) shows the energy of two of the parallel tempering chains; room temperature and 300°C. For real proteins, which have more complicated energy landscapes and possibly high energy barriers, it is difficult to know the temperature range required for parallel tempering to explore the entire parameter space and not be trapped in a particular basin. Nested sampling, with its top-down, temperature-independent approach, does not suffer from this problem.

Another of the advantages of nested sampling is that simulations are temperature-independent, and hence can provide estimates of thermodynamic variables at any temperature. Fig. 2 shows the heat capacity (C_v) curve for the 16-residue polyalanine. The curves were calculated using nested sampling (converged down to -25°C, so that the C_v curve does not stop abruptly at room temperature), and parallel tempering. The solid line is calculated using 10 nested sampling simulations each with 1.3×10^9 MC moves. The dashed lines show twice the standard error. The parallel tempering curve shows the heat capacity using 10 parallel tempering simulations (again each with 1.3×10^9 MC moves) with error bars showing twice the standard error. For parallel tempering, the heat capacity is only calculated for discrete temperatures and a procedure such as Boltzmann reweighting (36) is needed to calculate the continuous curve.

There appears to be good agreement between the methods. Previous results have found nested sampling to be more efficient at calculating the heat capacity curves (23). In this example, we found nested sampling to be of similar efficiency to parallel tempering. We believe this to be because, unlike the system presented in Pártay et al. (23), our phase transition (from coil to hairpin) occurs over a very large energy (and hence temperature) range from which parallel tempering can successfully sample.

Protein G

Protein G is a 56-residue protein consisting of an antiparallel four-stranded β-sheet and an α-helix, with a β-Grasp

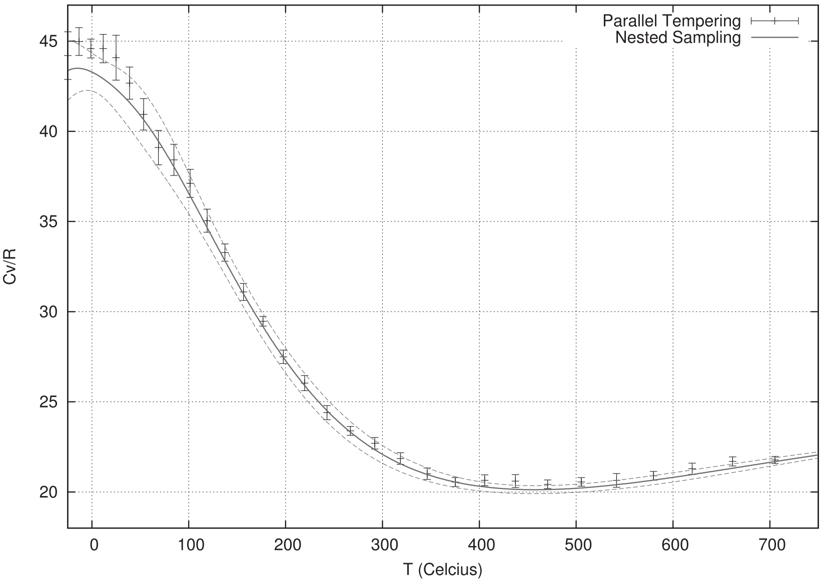


FIGURE 2 Heat capacity curve for the 16-residue polyalanine. The nested sampling simulations (*solid line*) use 1.3×10^9 MC moves, with error lines denoting two standard errors from the mean. The parallel tempering uses the same number of MC moves again with error bars showing two standard errors from the mean.

(ubiquitin-like) fold, which has been extensively studied by a variety of folding simulation techniques (37–40). Its native structure is shown on the left of Fig. S3. All figures of protein G in this article have been oriented so that the first β-strand is the second strand from the right and the N-terminal residue is at the top.

As described above, the nested sampling procedure can be used to estimate the thermodynamic energy of the system at any temperature, using Eq. 4 For protein G, at room temperature ($\beta = 1.0$), the thermodynamic energy is -190 units. Fig. S3 shows a sample of four room-temperature, thermodynamically accessible conformations found by a single nested sampling simulation with $K = 20{,}000$ and $m = 15{,}000$. The conformers have energies -189, -190, -191, and -190, respectively, with backbone RMSDs (from the crystal structure) of 1.93 Å, 2.96 Å, 3.97 Å, and 5.22 Å, respectively. The estimated value of the backbone RMSD at $\beta = 1$, calculated using Eq. 5, is $\mathbb{E}(\text{RMSD}|\beta = 1) = 3.21$ Å.

Conformers *A–D* in Fig. S3 have the correct backbone topology, close to the native structure, but there is a reasonable amount of variation in the orientation of the helix with respect to the β-sheet at this temperature. It is important to remember that protein structures are intrinsically flexible (41–43), and the crystal structure (1PGA.pdb) is only one member of an ensemble of conformations that the protein may explore. In the Supporting Material we demonstrate that flexible motion of protein G allows a substantial reorientation of the axis of the helix with respect to the sheet. Conformers *A–D* in Fig. S3, which differ from the native state principally in the orientation of the helix relative to the sheet, may therefore be more representative of the native state than the RMSD alone suggests.

The first half of the nested sampling simulation is spent exploring high-energy conformations with no noticeable secondary structure and often steric hindrances. In the second half of the simulation, once the long-range quadratic bias potential has pulled the secondary structure elements close together, the short-range hydrogen-bond interaction contributions increase to dominate the bias potential contributions, having a steeper gradient in the last third of the simulation (see Fig. S4 (*top*)). The short-range hydrophobic interaction contributions are the smallest, but nevertheless not negligible; they ensure the correct packing of the hydrophobic and amphipathic side chains at conformations available at room temperature (see below and Fig. 3). Fig. S4 (*bottom*) shows a sequence of 10 conformers in order of decreasing energy. These conformers come from the deepest basin of the energy landscape chart of a simulation (higher energy conformers come from the part of the energy landscape chart that contains the deepest basin). The sequence illustrates how the secondary and tertiary structure accrete in the course of a simulation, capturing the essence of the hierarchical folding model. The sequence is not, however, a single folding pathway, in the sense of a molecular dynamics trajectory; there are many conformations in the

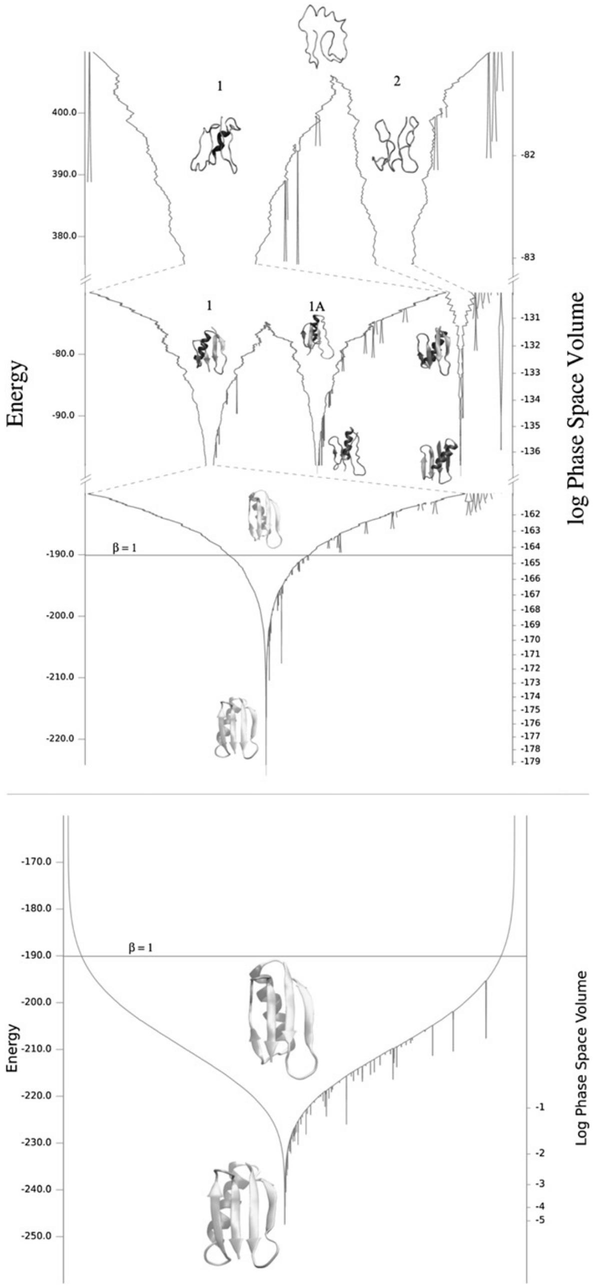


FIGURE 3 (*Top*) Prior (potential) energy landscape chart. (*Bottom*) Posterior energy landscape chart at $\beta = 1$, for a nested sampling simulation of protein G using $K = 20{,}000$ and $m = 15{,}000$. (*Left axis*) The energy is shown in units of *RT*, and the width of the chart is proportional to the sum of the prior (*top*) and posterior (*bottom*) weights of the nested sampling points below the given energy level (shown on the *right axis*). The prior energy landscape chart shows the potential energy surface and the posterior energy landscape chart shows, for a given temperature, the probabilities of finding conformations from the different basins. At $\beta = 1$ (room temperature), only funnel 1 is accessible. The scaling function used for the prior energy landscape chart is $\exp(-fE)$ (with f being 0.1, 0.4, and 0.7 on the *top*, *middle*, and *bottom panels* of the prior energy landscape chart, respectively). Example conformers from the main basins, at various energy levels, are shown on the charts.

active set in the same basin. It is possible, though, that Fig. S4 (*bottom*) represents a plausible sequence of events leading to the native structure.

Energy landscape charts using the prior and posterior weights for a nested sampling simulation of protein G using $K = 20{,}000$ and $m = 15{,}000$, calculated using a connectivity number $k = 15$, are shown in Fig. 3. The volume scale on the right-hand axis shows the proportion of the prior and posterior phase space volume available below the given energy level. The width of the chart uses this scale. Basins that contain <1/1000th of the probability mass at the point of splitting are not shown on the diagram. Conformers have been placed on the chart to provide examples of the samples found in different places of the chart.

Topologically, for energy above 405 units, there is one main basin containing virtually all of the samples. There is little structure in the samples, as shown by the conformer at the top of the chart. However, at energy 405 units, the phase space splits into two main funnels: one with the helix forming on the correct side of the sheet (funnel 1) and one with it forming on the incorrect side (funnel 2). Funnel 1 further splits at energy -75 units, corresponding to conformations where the hydrophobic residues are in the interior of the protein (funnel 1) or on the surface of the protein (funnel 1A). At room temperature (the expected energy corresponding to $\beta = 1$ is marked by a *horizontal line* on both panels of the chart), the phase space volume of both funnels 1A and 2 are <1/1000th of the main funnel and hence the posterior energy landscape chart consists of a single funnel. The inaccessibility of funnel 1A at room temperature indicates the importance of hydrophobic interactions.

Fig. S5 shows two conformers that are placed in the same small basin, branching off the right-hand funnel. The conformer on the left has higher energy than the one on the right. These conformers are very similar, and demonstrate that the topological analysis shows how metastable conformations are formed. The pathway to these states would be obtained by considering conformers found in the same basin.

DISCUSSION

It is interesting to consider how the energy landscape charts vary from simulation to simulation. Topologically, we always find two main funnels in the protein G simulations (funnels 1 and 2 on Fig. 3), corresponding to the packing of the helix on either side of the sheet. The dominant mode with the native like backbone topology (funnel 1) splits again at a lower energy level to two funnels, corresponding to the hydrophobic residues being in the interior (funnel 1) or on the surface (funnel 1A) of the protein. The energy at which funnels 1 and 2 split varies significantly between simulations, from 220 to 580 energy units. This is probably because the RMSD metric is an imprecise way of comparing wildly different conformations. The energy where funnels 1 and 1A split has a much smaller variation, -75 to -55 energy units. This trend in the variation of splitting energies was also observed in the nested sampling simulations of the other modeled proteins. Metrics other than the RMSD might improve the reproducibility of energy landscape charts and would be worthy of investigation.

The relative basin widths of energy landscape charts depend on the size of the nested sampling active set, K. In general, K determines the resolution of exploration. When converging the evidence at lower temperatures, a larger value of K is required. This is because at every splitting of the likelihood function, the probability of exploring the dominant mode decreases, according to Eq. 3. At high energies, the accessible conformational space is connected, and the MMC procedure explores the space effectively. As the energy lowers, the accessible conformational space becomes increasingly disconnected. Because the MMC procedure cannot jump between disconnected components of the conformational space, an increasingly large set of active points is required to sample effectively. As the posterior mass is concentrated at lower energies for lower temperatures, K behaves as an effective minimum temperature. Using too small an active set for a given temperature causes large variation between different nested sampling simulations; for example, the estimates for the evidence and the relative widths of the funnels of energy landscape charts.

In the protein G simulations, we find that $K = 20{,}000$ is large enough to produce simulation independent charts for temperatures near $\beta = 1$. When using, for example, $K = 2500$, which is too small for sampling the posterior distribution at $\beta = 1$, we find that the active set becomes extremely homogenous and the simulation is, in effect, exploring just one tiny basin in one of the main funnels, by making smaller and smaller crankshaft rotations. Hence, we find a single room-temperature accessible conformation, as opposed to the wide selection that is found when $K = 20{,}000$.

The magnitude of m relative to K is problem-specific. It has been suggested that for probability distributions that lack a large number of modes, it is optimal to set K small and use a large m (the cost is proportional to mK) (21). For protein G, we find the energy landscape is so complex that we need a large K to explore all the funnels simultaneously, and a large m to ensure the active set remains heterogeneous, and we therefore choose m and K to have the same order of magnitude. Incorporating nonlocal flexible motions (44) into our MMC procedure may allow a decrease in m without losing heterogeneity and this is a focus of future work. If this proves to be the case, we would choose to increase K relative to m.

In our previous work (27), using MMC with parallel tempering to simulate the folding of protein G with a simpler model (no γ-atoms and hydrophobic interactions were included in this model), the lowest energy structures obtained were similar to those shown at the bottom of funnel 2 of Fig. 3, with the helix packed on the incorrect side of the sheet and a backbone RMSD of 8.6 Å from the crystal structure. This demonstrates the difficulty of using parallel tempering or simulated annealing to reconstruct the native structure, when the energy landscape exhibits two main funnels separated by a large energy barrier. If the annealing proceeds down the incorrect funnel it will be nearly impossible for it to climb back out and down into the correct funnel.

The reason for the double funnel is the symmetry of the protein G topology with respect to the Gō-type bias potential, which is the predominant factor at the beginning of the simulation. The further splitting of the main funnel into funnels 1 and 1A (Fig. 3) is also due to the nature of the Gō-type bias potential. This applies a quadratic potential on the C_β atom contacts, which does not restrict the hydrogen-bond pattern between the individual strands; at high energies, both conformations (with the hydrophobic residues of the β-sheet being in the interior or on the surface of the protein) are similarly likely to be adopted. However, other energy and entropy contributions due to the presence of side chains (e.g., hydrophobic interactions and steric clashes) ensure that only conformations with the nativelike topology are accessible at room temperature. This way, energy landscape charts also reflect the nature of the protein model and force field used. For example, the energy landscape charts for chymotrypsin inhibitor 2, which differs in topology from protein G, but also possesses a similar symmetry with regard to the packing of the α-helix against the β-sheet, also exhibit this double funnel (see the Supporting Material).

It would be interesting to compare energy landscape charts of nested sampling simulations using other protein models and force fields, for example, all-atom representations, and this will be a focus of future work.

CONCLUSION

This article has described the parallelization of the nested sampling algorithm, and its application to the problem of protein folding in a force field of empirical potentials that were designed to stabilize secondary structure elements in room-temperature simulations. The output of the nested sampling algorithm can be used to produce energy landscape charts, which give a high level description of the potential energy surface for the protein folding simulations. These charts provide qualitative insights into both the folding process and the nature of the model and force field used. The topology of the protein molecule emerges as a major determinant of the shape of the energy landscape, as has been noted by other authors (37). The energy landscape chart for protein G exhibits a double funnel with a large energy barrier, a potential energy surface that parallel tempering struggles to explore fully. The nested sampling algorithm also provides an efficient way to calculate free energies and the expectation value of thermodynamic observables at any temperature, through a simple postprocessing of the output.

SUPPORTING MATERIAL

Supporting text with 11 figures and references (45–62) is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)00055-0.

We thank Gábor Csányi, Konrad Dabrowski, Lívia Pártay, Alexei Podtelezhnikov, and John Skilling for helpful discussions.

We acknowledge support from the Leverhulme Trust (grant F/00 215/BL to N.S.B., C.V., and D.L.W.) and the Engineering and Physical Sciences Research Council, UK (grant EP/GO21163/1 to D.L.W.). S.A.W. acknowledges support from the Leverhulme Trust (Early Career Fellowship).

REFERENCES

1. Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science.* 181:223–230.
2. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA.* 84:7524–7528.
3. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
4. Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24:26–33.
5. Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* 24:77–83.
6. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2011. How fast-folding proteins fold. *Science.* 334:517–520.
7. Gō, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.
8. Takada, S. 1999. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA.* 96:11698–11700.
9. Dinner, A. R., A. Sali, …, M. Karplus. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25:331–339.
10. Sali, A., E. Shakhnovich, and M. Karplus. 1994. How does a protein fold? *Nature.* 369:248–251.
11. Clementi, C., A. E. García, and J. N. Onuchic. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.* 326:933–954.
12. Hori, N., G. Chikenji, …, S. Takada. 2009. Folding energy landscape and network dynamics of small globular proteins. *Proc. Natl. Acad. Sci. USA.* 106:73–78.
13. Amato, N. M., K. A. Dill, and G. Song. 2003. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.* 10:239–255.
14. Becker, O. M., and M. Karplus. 1997. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* 106:1495–1517.
15. Wales, D. J. 2004. Energy Landscapes: Applications to Clusters, Biomolecules and Glasses. Cambridge University Press, Cambridge, UK.
16. Wales, D. J., and T. V. Bogdan. 2006. Potential energy and free energy landscapes. *J. Phys. Chem. B.* 110:20765–20776.
17. Skilling, J. 2006. Nested sampling for general Bayesian computation. *J. Bayesian Anal.* 1:833–860.
18. Sivia, D. S., and J. Skilling. 2006. Data Analysis, A Bayesian Tutorial, 2nd ed. Oxford University Press, Cambridge, UK.

19. Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671–680.

20. Wang, F., and D. P. Landau. 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86:2050–2053.

21. Swendsen, R. H., and J. S. Wang. 1986. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* 57:2607–2609.

22. Murray, I., D. MacKay, …, J. Skilling. 2006. Nested sampling for Potts models. *Adv. Neural Inf. Process. Syst.* 18:947–954.

23. Pártay, L. B., A. P. Bartók, and G. Csányi. 2010. Efficient sampling of atomic configurational spaces. *J. Phys. Chem. B.* 114:10502–10512.

24. Bennett, C. H. 1976. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22:245–268.

25. Mukherjee, P., D. Parkinson, and A. R. Liddle. 2006. A nested sampling algorithm for cosmological model selection. *J. Astrophys Lett.* 638:51–54.

26. MacKay, D. J. C. 2004. www.inference.phy.cam.ac.uk/bayesys/box/nested.pdf.

27. Podtelezhnikov, A. A., and D. L. Wild. 2009. Reconstruction and stability of secondary structure elements in the context of protein structure prediction. *Biophys. J.* 96:4399–4408.

28. Podtelezhnikov, A. A., and D. L. Wild. 2008. CRANKITE: a fast polypeptide backbone conformation sampler. *Source Code Biol. Med.* 3:12.

29. Podtelezhnikov, A. A., Z. Ghahramani, and D. L. Wild. 2007. Learning about protein hydrogen bonding by minimizing contrastive divergence. *Proteins.* 66:588–599.

30. Podtelezhnikov, A. A., and D. L. Wild. 2005. Exhaustive Metropolis Monte Carlo sampling and analysis of polyalanine conformations adopted under the influence of hydrogen bonds. *Proteins.* 61:94–104.

31. Srinivasan, R., and G. D. Rose. 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci. USA.* 96:14258–14263.

32. Creamer, T. P., and G. D. Rose. 1995. Interactions between hydrophobic side chains within α-helices. *Protein Sci.* 4:1305–1314.

33. Luom, P., and R. L. Baldwin. 2002. Origin of the different strengths of the $(I,_{i+4})$ and $(I,_{i+3})$ leucine pair interactions in helices. *J. Biophys. Chem.* 96:103–108.

34. Elofsson, A., S. M. Le Grand, and D. Eisenberg. 1995. Local moves: an efficient algorithm for simulation of protein folding. *Proteins.* 23: 73–82.

35. Sibanda, B. L., T. L. Blundell, and J. M. Thornton. 1989. Conformation of β-hairpins in protein structures. A systematic classification with applications to modeling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* 206:759–777.

36. Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo analysis. *Phys. Rev. Lett.* 63:1195–1198.

37. Karanicolas, J., and C. L. Brooks, 3rd. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.

38. Kolinski, A., and J. Skolnick. 2004. Reduced models of proteins and their applications. *Polymer (Guildf.).* 45:511–524.

39. Sheinerman, F. B., and C. L. Brooks, 3rd. 1998. Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* 278:439–456.

40. Shimada, J., and E. I. Shakhnovich. 2002. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. USA.* 99:11175–11180.

41. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. *Nature.* 450:964–972.

42. Thorpe, M., M. Lei, …, L. A. Kuhn. 2001. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.* 19:60–69.

43. Wells, S., S. Menor, …, M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* 2:S127–S136.

44. Jimenez-Roldan, J. E., R. B. Freedman, …, S. A. Wells. 2012. Protein flexibility explored with normal modes and geometric simulation. *Phys. Biol.* In press.

45. Engh, R. A., and R. Huber. 1991. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallogr. A.* 47:392–400.

46. Engh, R. A., and R. Huber. 2001. Structure quality and target parameters. *In* International Tables for Crystallography, Vol. F: Crystallography of Biological Macromolecules, 1st ed. M. G. Rossman and E. Arnold, editors. Kluwer Academic Publishers for the International Union of Crystallography, Dordrecht, Boston, London. 382–392.

47. Brünger, A. 1992. X-PLOR, V. 3.1: A System for X-Ray Crystallography and NMR. Yale University Press, New Haven, CT.

48. Ho, B. K., E. A. Coutsias, …, K. A. Dill. 2005. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* 14:1011–1018.

49. Brenner, S. E., P. Koehl, and M. Levitt. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 28:254–256.

50. Carreira-Perpinan, M., and G. Hinton. 2005. On contrastive divergence learning. *In* Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Jan. 6–8, 2005, The Savannah Hotel, Barbados. 217.

51. Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7:95–99.

52. Hopfinger, A. J. 1973. Conformational Properties of Macromolecules. Academic Press, New York.

53. Word, J. M., S. C. Lovell, …, D. C. Richardson. 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* 285:1711–1733.

54. Pappu, R. V., R. Srinivasan, and G. D. Rose. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA.* 97:12565–12570.

55. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein DataBank. *Nucleic Acids Res.* 28:235–242.

56. Baker, E. N., and R. E. Hubbard. 1984. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* 44:97–179.

57. Savage, H. J., C. J. Elliott, …, J. M. Finney. 1993. Lost hydrogen bonds and buried surface area: rationalizing stability in globular proteins. *J. Chem. Soc., Faraday Trans.* 89:2609–2617.

58. Stickle, D. F., L. G. Presta, …, G. D. Rose. 1992. Hydrogen bonding in globular proteins. *J. Mol. Biol.* 226:1143–1159.

59. McDonald, I. K., and J. M. Thornton. 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238:777–793.

60. Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

61. Word, J. M., S. C. Lovell, …, D. C. Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285:1735–1747.

62. Suhre, K., and Y.-H. Sanejouand. 2004. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* 32(Web Server issue):W610–W614.

Supporting Material

Figures Referenced in Main Text

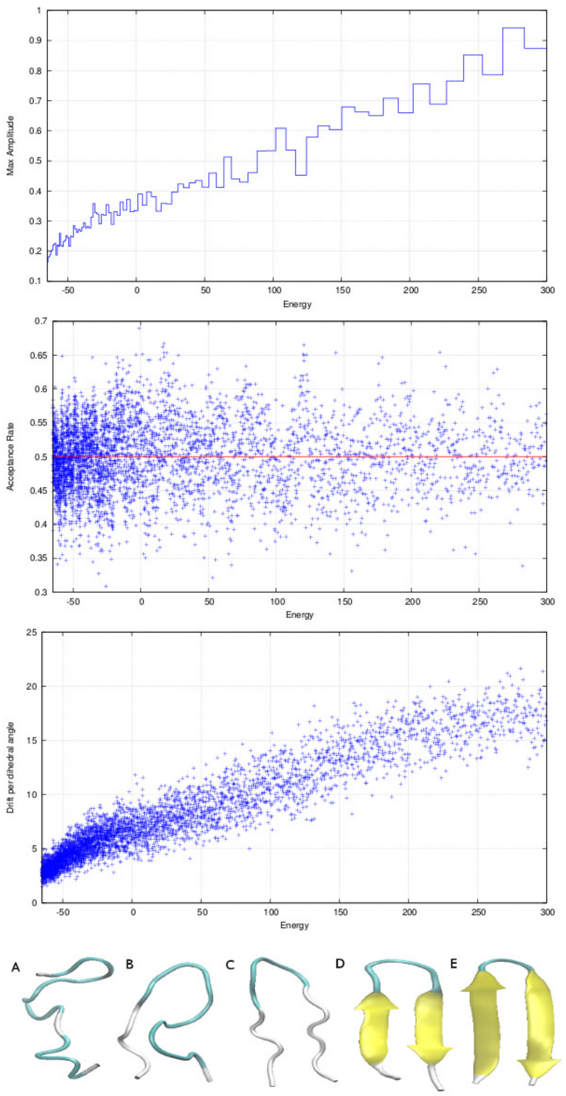


Figure S1: Top graph: the maximum allowed amplitude of the crankshaft rotations α_0 (in radians); Middle graph: the acceptance rate of the MC chains; Bottom graph: the drift per dihedral angle (the distance between the start and end conformations of a single MC chain). All with respect to the current energy threshold, for a 16 residue polyalanine β-hairpin; see text for more details. Bottom: Five snapshots from a single nested sampling simulation of a β-hairpin with $K = 1000$ and $m = 2500$. The snapshots are equally spaced along the $\log(X)$ axis and have energies 3567, 190, 0, -46 and -66 (A-E). For comparison, the expectation of the internal energy at room temperature is -43.

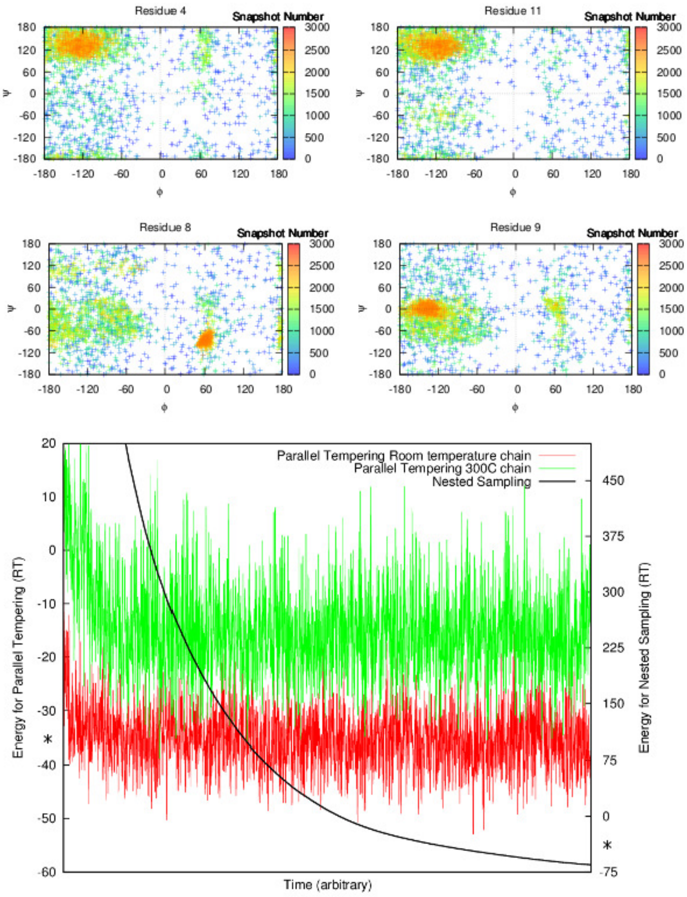


Figure S2: Top: Dihedral angle evolution for residues 4, 8, 9 and 11 of the 16 residue polyalanine nested sampling simulation. In the later snapshots, residues 4 and 11 are distributed in the standard β-sheet region of the Ramachandran plot. Residues 8 and 9 contain the turn of the polypeptide. The dihedral angles of the turn residues, 8 $(60 \pm 15, -90 \pm 30)$ and 9 $(-150 \pm 30, 0 \pm 30)$, are closest to the values of type II' turn $((60, -120)$ and $(-80, 0))$ Bottom: Energy v Time graph for nested sampling (right hand axis) and two of the chains from a parallel tempering simulation (room temperature and 300°C; both left hand axis.). On both vertical axes a star marks the expected thermodynamic energy at room temperature.

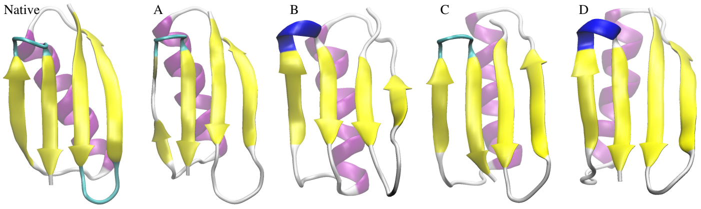


Figure S3: The native (crystal) structure of protein G (left) with a sample of conformations accessible at room temperature from a simulation with $K = 20000$ and $m = 15000$. All figures of protein G in this paper have been oriented so that the first β-strand is the second strand from the right and the N-terminal residue is at the top. The thermodynamic energy at room temperature, estimated from the simulation, is -190 and conformers A,B,C and D have have energies -189, -190, -191 and -190, respectively. The backbone RMSDs from the crystal structure are 1.93 Å, 2.96 Å, 3.97 Å and 5.22 Å, respectively. The angle between the helix projected onto the sheet and the first β-strand is 17.9°, 8.6°, $-4.7°$ and $-15.1°$, respectively, compared to 21.9° of the crystal structure.

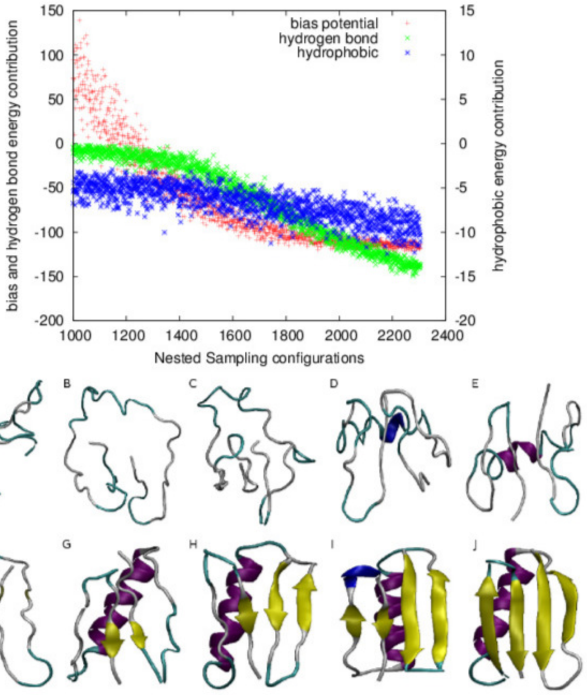


Figure S4: Top: Energy contributions of the Gō-type bias potential (red), hydrogen bonds (green) and hydrophobic interactions (blue) in the second half of a nested sampling simulation of protein G with $K = 20000$ and $m = 15000$. Units of energy are in RT corresponding to temperature. Note the different scale on the vertical axes. Bottom: Ten conformations of protein G from the same simulation, in order of decreasing energy.

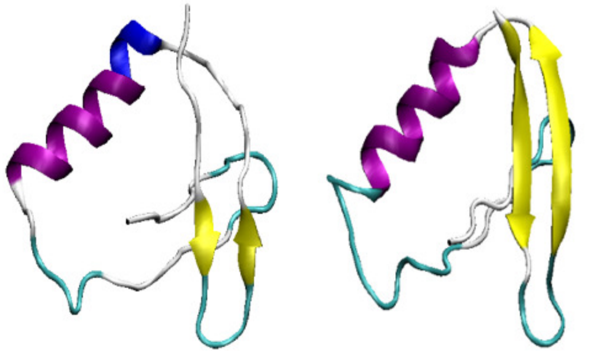


Figure S5: The topological analysis places similar conformations in the same basin. For example, these two conformations (which both have the 3rd and 4th β-strands aligned incorrectly) are placed in the same basin.

1 The Protein Model

We modelled the polypeptide as a chain of peptide groups elastically connected at the C_α atoms, with the valence angles constrained to $111.5° \pm 2.8°$. The positions of all backbone and C_β atoms, including hydrogen, were specified by the orientations of the peptide bonds. We fixed the peptide bond lengths and angles at standard values (1–3). The distance between C_α atoms separated by *trans* peptide bonds was fixed at 3.8 Å. The C_β positions were stipulated by the tetrahedral geometry of the C_α atoms and corresponded to L-amino acids. Most of the conformational variability of polypeptides comes from relatively free rotation around N–C_α and C_α–C bonds characterised, respectively, by dihedral angles ϕ and ψ (Fig. 1 in (4)). These rotations are least restricted in glycine that lacks C_β. The dihedral angles ϕ in proline were elastically constrained to $-60° \pm 7°$ by covalent bonding (5). We introduced a harmonic potential E_i^B to impose these and other elastic constraints. A more detailed description of the model is given in our previous work (4).

In this work, we represented other side chain atoms by one, or in the case of branched side chains, two pseudo-atoms, following (6). The side chain dihedral angles χ were permitted to vary, and take the values $\{\pm 60°, 180°\}$, or in the case of proline $\{\pm 30°\}$, with probabilities dependent on residue type, with values corresponding to the distribution of the χ angles in the same ASTRAL PDB database (7) that was used in (4), and here, to learn the potential parameters by a statistical machine learning procedure, contrastive divergence (8).

We modelled van der Waals repulsions so that there is a prohibitively large energetic cost of overlaps between atoms. We used values of atomic radii close to a lower limit of the range found in the literature (9–12): $r(C_\alpha) = r(C_\beta) = 1.57$ Å, $r(C) = 1.42$ Å, $r(O) = 1.29$ Å, $r(N) = 1.29$ Å. We adopted values of the contact radii for the pseudo-atoms from (6).

Hydrogen bonding is a major polar interaction between NH and CO groups of polypeptide backbone. Based on surveys of the Protein Data Bank (PDB) (13), important reviews of hydrogen bonding in globular proteins have formulated the basics of the current understanding of hydrogen bond geometry and networking (14–17). We considered the hydrogen bond formed when three distance and angular conditions were satisfied: $r(\mathrm{O, H}) < \delta$, $\angle\mathrm{OHN} > \Theta$, and $\angle\mathrm{COH} > \Psi$, where $r(\mathrm{O, H})$ is the distance between oxygen and hydrogen, and symbol $\angle$ denotes the angle between the three atoms (see Fig. 1A in (18)). The lower bound on the separation between the atoms ($r(\mathrm{O, H}) > 1.8$ Å) was implicitly set by the hard-sphere collision between oxygen and nitrogen. We used the same hydrogen bond potential regardless of the secondary structure adopted by the peptide backbone. The energy of the hydrogen bond (Fig. 1B in (18)) was described in (18) by a square-well potential,

$$E_{ij}^{HB} = -n_h H \tag{1}$$

where H is the strength of each hydrogen bond, and n_h is the number of hydrogen bonds between the amino acids i and j. The strength of the hydrogen bonds, H, as well as the three cutoff parameters, δ, Θ, Ψ was determined by a machine learning procedure, contrastive divergence (8). We found that softening the hard cutoffs $\{\delta, \Psi, \Theta\}$ improved the results, and hence we used a steep continuous approximation to the square well.

We modelled hydrophobic interactions in a manner consistent with (6). The hydrophobic interaction contribution between hydrophobic atoms A and B of amino acids i and j ($|i - j| \geq 2$) is

$$E_{AB}^{hyd} = \begin{cases} f k_h & r_{AB} < r_{cut,AB} \\ \\ f k_h \left(\frac{r_{AB} - r_{cut,AB}}{\Delta}\right) & r_{cut,AB} \leq r_{AB} < r_{cut,AB} + \Delta \end{cases} \tag{2}$$

where k_h is a constant parameter proportional to the Kauzmann parameter (19), the cutoff distance $r_{cut,AB}$ is the sum of the vdW radii of atoms A and B listed in (6), $\Delta = 2.8$ Å is a smoothing range beyond the cutoff distance, and the multiplicative factor f takes the value 2, if both amino acids are hydrophobic, 1, if one is hydrophobic and the other one is amphipathic, and 0, if neither are hydrophobic. Hydrophobic amino acids are cysteine, isoleucine, leucine, methionine, phenylalaninne, tryptophan and valine; amphipathic residues are alanine, histidine, threonine and tyrosine.

The sequence-*dependent* part of the potential (the negative log-likelihood) was approximated in our model by pair-wise interactions between side-chains, as described in (20). Our main focus was on the resulting effect of these interactions and how they stabilise secondary structural elements. We did not consider the detailed physical nature of these forces, or how they depend on the amino acid types. We introduced these interactions between the polypeptide side chains as an effective Gō-type potential (21) dependent on the distance between C_β atoms,

$$E_{ij}^{SC} = \kappa C_{ij}(r_{ij} - r)^2 \tag{3}$$

where r_{ij} is a distance between non-adjacent C_β atoms, $|i - j| > 1$; r a constant and κ is a force constant. In (20) we introduced a "regularised contact map", C_{ij}. In this binary matrix, two types of contacts were defined in the context of protein secondary structure. First, only lateral contacts in the parallel and anti-parallel β-sheets were indicated by 1's. Second, the contacts between amino acids i and $i + 3$ in α-helices were also represented by 1's. These contacts typically have the closest C_β–C_β distance among non-adjacent contacts in native proteins. The force constants and r depend on the secondary structure type, introducing positive κ_α κ_β, r_α and r_β. Non-adjacent contacts in secondary structural elements were, therefore, stabilised by attracting potentials.

We also modelled interactions between sequential residues. This interaction was defined by the mutual orientation of adjacent residues that are involved in secondary structural elements,

$$E_{i,i+1}^{SC} = \eta \cos \gamma_{i,i+1} \tag{4}$$

where $\gamma_{i,i+1}$ is the dihedral angle N_i–$C_{\alpha,i}$–$C_{\alpha,i+1}$–C_{i+1} between the adjacent residues. The purpose of this interaction is to bias the conformation towards the naturally occurring orientations of residues in secondary structural elements. In α-helices, adjacent residues adopt a conformation with positive $\cos \gamma$. In β-sheets, $\cos \gamma$ is negative. We, therefore, used two values of the force constant: negative η_α and positive η_β.

As in (20), all parameters were determined by a statistical machine learning procedure, contrastive divergence (8) and in this work $\delta = 2.06$, $-\cos\Theta = 0.89$, $-\cos\Psi = 0.766$, $H = 4.35$, $\eta_\beta = 3.5$, $\eta_\alpha = -4.9$, $\kappa_\alpha = 3.3$, $\kappa_\beta = 1.2$, $r_\alpha = 5.66$, $r_\beta = 5.35$ and $k_h = 0.08$, where the unit of energy is RT at room temperature. With the improved model and force field described in this paper, contrastive divergence provided good parameters without the need of further adjustments, as had been the case in (20).

To summarise, the total energy of a polypeptide chain with conformation Ω was calculated as follows

$$E(R, \Omega) = \sum_{i=1}^{N} E_i^B + \sum_{i=1}^{N}\sum_{j=1}^{i}(E_{ij}^{vdW} + E_{ij}^{HB} + E_{ij}^{SC} + E_{ij}^{hyd}) \tag{5}$$

where we consider harmonic valence elasticity, E_i^B, van der Waals repulsions, E_{ij}^{vdW}, hydrogen bonding, E_{ij}^{HB} and hydrophobic packing, E_{ij}^{hyd}. The valence elasticity, van der Waals repulsions, and hydrogen bonding that contribute to this potential have a clear physical meaning and are analogous to traditional *ab initio* approaches. The side-chain interactions, E_{ij}^{SC} in this model were introduced as a long-range quadratic Gō-type potential based on the contact map and secondary structure assignment. This pseudo-potential had two purposes: it was needed to stabilise the secondary structural elements, and to provide a biasing force that allows reconstruction of the backbone conformation in the course of Metropolis Monte Carlo simulations (4, 20).

2 Additional Results

Src Tyrosine Kinase SH3 Domain.

Src Tyrosine Kinase SH3 Domain is a 56-residue protein, comprising a 5-stranded β-barrel. The last strand is interrupted by a single turn of a 3_{10}-helix, which was not included in the 'regularised' contact map used to define the native state in these simulations. The native (crystal) structure is shown in Fig. S6.

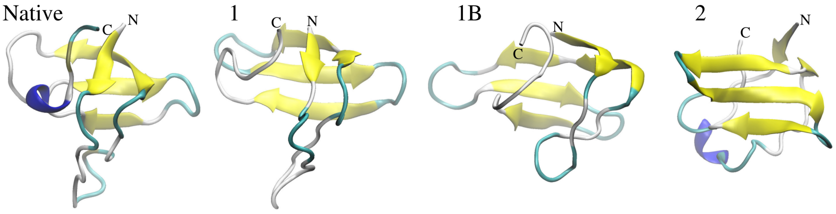


Figure S6: The native (crystal) structure of the SH3 Domain, with conformers from the two main funnels accessible at room temperature from a simulation with $K = 15000$ and $m = 15000$. Conformers 1 and 1B have the correct backbone topology, unlike conformer 2. The backbone RMSDs are 4.89 Å, 5.28 Å and 11.51 Å, respectively. The N and C termini are marked; see the text for more details.

The energy landscape for a simulation with $K = 15000$ and $m = 15000$, depicted in Fig. S7, shows two main funnels (funnels 1 and 2) that further split into sub-funnels (funnels 1 and 1A, and funnels 2 and 2A). At room temperature, conformations representative of funnels 1 and 2 are accessible, and most of the posterior mass is in funnel 1, which contains conformers with the correct backbone topology. Funnel 1 splits further at −43 energy units. The sub-funnels 1 and 1B are connected at room temperature, and the probability of moving from one to the other is non-zero, due to the non-zero probability of adopting a conformation that is indistinguishable from those typical to funnels 1 and 1B. The estimated energy at room temperature is −39 energy units.

Fig. S6 shows low energy conformers from the same simulation. Conformers 1 and 1B (from funnels 1 and 1B in Fig. S7) have the correct backbone topology and a backbone RMSD of 4.61 Å and 5.47 Å with respect to the crystal structure. Moreover, although the N-terminal loop is not included in the 'regularised' contact map used in the simulation, the packing of the loop is in reasonable agreement with the crystal structure. Conformer 2, taken from the bottom of the other major funnel, adopts a conformation with an incorrect backbone topology – note the relative positions of the N and C termini with respect to the sheet. The conformers shown in Fig. S6 correspond to lower energies than the estimated energy at room temperature, but have been shown as it is clearer to see the differences between them once the β-strands have fully formed. Conformations available at room temperature typically have shorter β-strands. The estimated backbone RMSD at room temperature is $\mathbb{E}(\text{RMSD}|\beta = 1) = 6.46$ Å.

Chymotrypsin Inhibitor 2.

Chymotrypsin inhibitor 2 is a 65-residue protein which contains a four-stranded β-sheet and an α-helix, but differs in topology from protein G. However, it also possesses a similar symmetry with regard to the packing of the α-helix against the β-sheet.

As with protein G, the energy landscape shows two main folding funnels. Fig. S9 shows the prior energy landscape chart and the posterior energy landscape chart at room temperature for a simulation with $K = 15000$ and $m = 15000$. Sample conformers from the main funnels at different energy levels are also included. The posterior energy landscape chart shows that virtually all the posterior mass is in funnel 1 (including funnel 1A) at room temperature, with funnel 2 being insignificant. Funnel 1 splits at −75 energy units into funnels 1 and 1A, which are connected at room temperature; the probability of adopting a conformation that is indistinguishable from the ones in funnels 1 and 1A is non-zero. The estimated value of the energy at $\beta = 1$ is −82 units.

The native structure and sample conformers from the main funnels are also compared in Fig. S8. As with protein G, in our model, the helix can be packed on either side of the sheet, and in Fig. S8 conformers 1 and 2 are taken from the bottom of the two funnels. Conformer 1 has the correct backbone topology, whereas conformer 2 has the helix packed on the incorrect side of the sheet. Conformer 1A from funnel 1A has the correct backbone topology, but the hydrophobic residues of the C-terminal β-strand are packed on the wrong

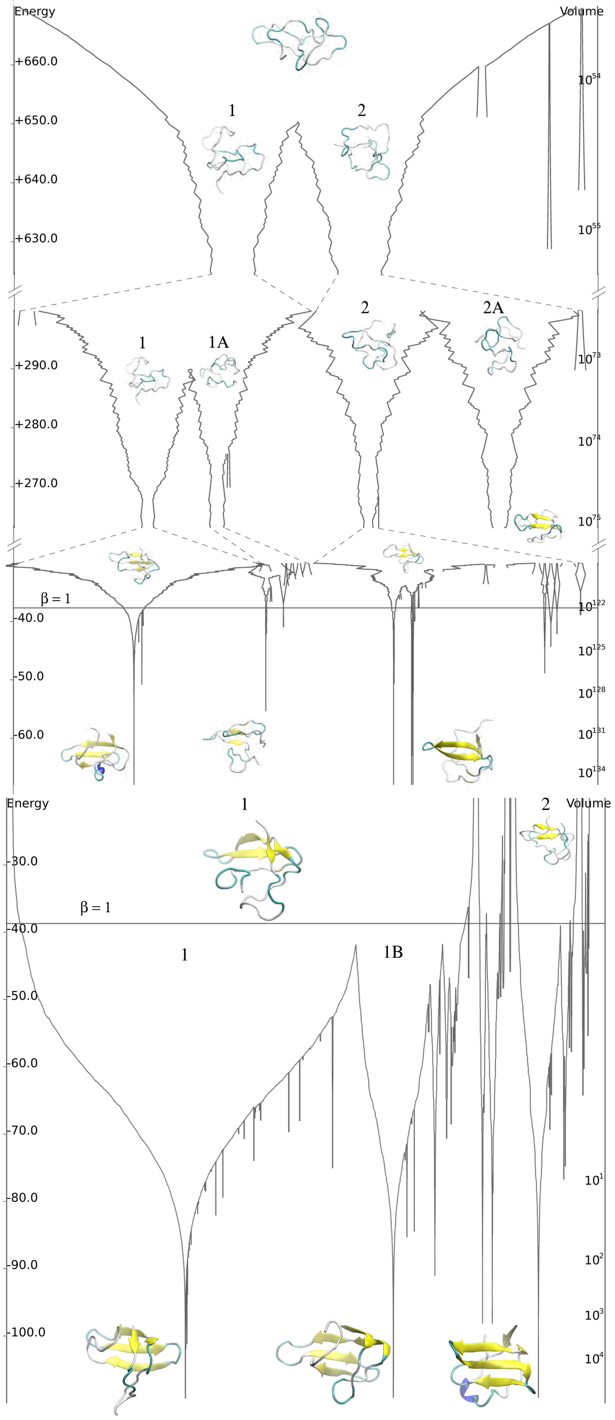


Figure S7: Top: prior (potential) energy landscape chart, Bottom: posterior energy landscape chart at $\beta = 1$ for a simulation of the SH3 Domain using $K = 15000$ and $m = 15000$. Sample conformers of funnels 1, 1A, 1B, 2 and 2A are marked on the chart.

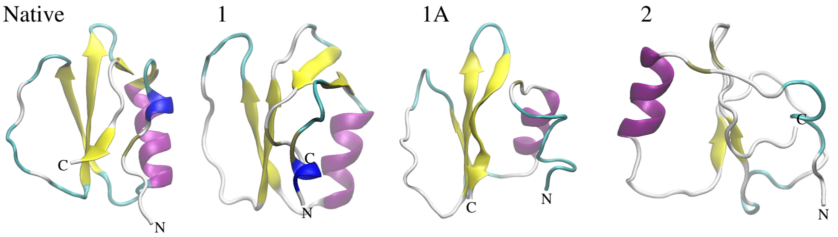


Figure S8: The native (crystal) structure for chymotrypsin inhibitor 2 (left) and 3 conformers from a simulation with $K = 15000$ and $m = 15000$. 1 and 1A: conformers from near the bottom of the funnel which contains conformers with the correct topology with backbone RMSD of 4.86 Å and 4.91 Å, respectively. The 1A conformer has the hydrophobic residues of the C-terminal β-strand on the wrong side. 2: a conformer from near the bottom of the other funnel with backbone RMSD of 11.18 Å. Note that the helix is packed on the wrong side of the sheet in conformer 2. The N and C termini are shown.

side of the sheet. Both conformers 1 and 1A have much lower energy than found at room temperature, and it is interesting to note that in both of these conformers the C-terminal β-strand has formed spontaneously without contact bias. In the crystal structure this strand is actually a large coil. In simulations, we find that the secondary structure that is defined by the regularised contact map forms first, and, since the model allows a large amount of freedom for residues which do not have contact bias, nested sampling then tries to place the remaining residues in the lowest energy position possible. The backbone RMSD of conformers 1, 1A and 2, from the crystal structure, are 4.86 Å, 4.91 Å and 11.18 Å, respectively, while the estimated backbone RMSD at room temperature is $\mathbb{E}(\text{RMSD}|\beta = 1) = 5.55$ Å.

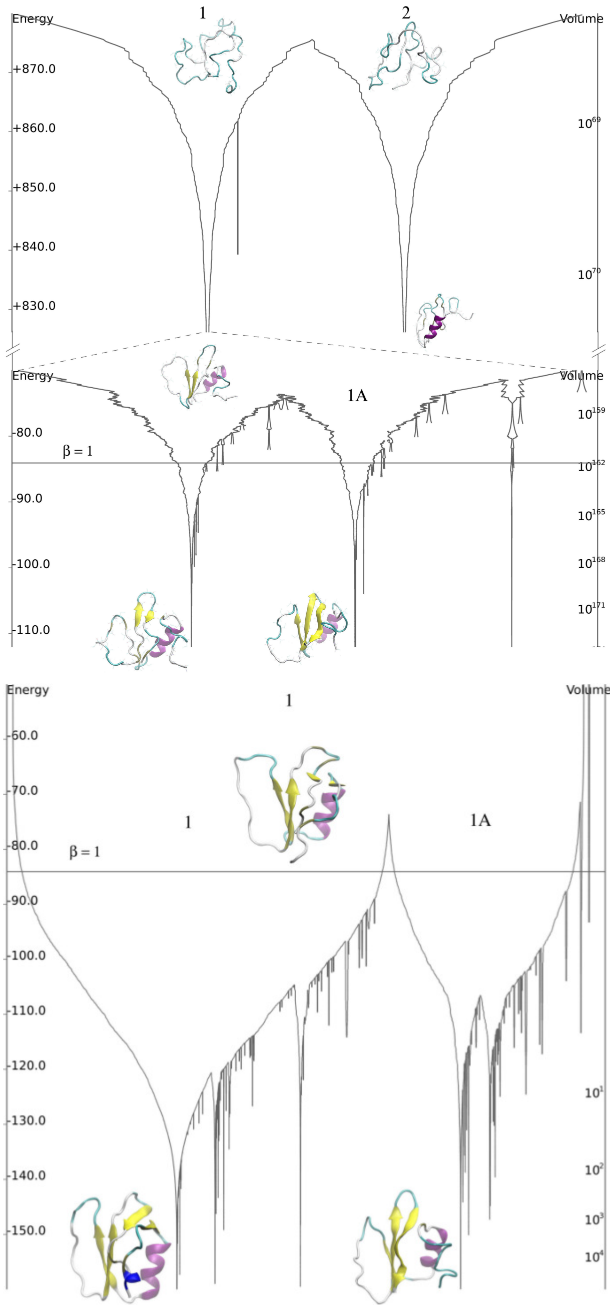


Figure S9: Top: prior (potential) energy landscape chart, Bottom: posterior energy landscape chart at $\beta = 1$ for a simulation of the chymotrypsin inhibitor 2 using $K = 15000$ and $m = 15000$. Sample conformers of funnels 1, 1A and 2 are marked on the charts.

3 Rigidity and Flexibility of Protein G

We have carried out rigidity analysis on Protein G using the software "FIRST" (22). Our input was the 1PGA.pdb structure with hydrogens added using the "Reduce" software (23). Rigidity analysis balances the degrees of freedom of the atoms against the constraints introduced by covalent bonding, hydrophobic tethers, salt bridges and hydrogen bonds. The result is a decomposition of the structure into rigid and flexible regions, known as a rigid cluster decomposition. The set of hydrogen bonds included is determined by a (negative) cutoff energy E_{cut}. An E_{cut} value near zero will include a large number of weak bonds and largely rigidify a structure; progressively lowering E_{cut} eliminates hydrogen bonds in an order from weakest to strongest, a process known as "rigidity dilution". The progress of this loss of rigidity can be mapped in a dilution plot (Figure S10a) in which the rigid cluster membership of each residue is mapped onto a 1-D representation of the protein backbone. A new line is plotted for each cutoff energy at which the rigidity of the mainchain changes.

At cutoff energies above -1.844 kcal/mol, the helix and the beta-sheet form a single rigid cluster (Figure S10b), while at lower energies the helix is a rigid body but the beta-sheet has become flexible (Fig. S10c). We stress that the backbone–backbone hydrogen bonding in both the helix and the sheet persists to much lower cutoff values. Once the helix and the sheet are not a single rigid cluster, motion of the helix with respect to the sheet becomes possible. The amplitude of such motion will be constrained by covalent and non-covalent interactions, in particular the many hydrophobic tethers between helix and sheet residues.

We obtain an eigenvector for flexible motion using a coarse-grained (one site per residue) elastic network model as implemented in the software "ElNeMo" (24). The lowest-frequency non-trivial mode, mode 7, corresponds to a rotation of the helix about an axis perpendicular to the beta-sheet. Linear projection of the structure along this mode would rapidly introduce unphysical distortions such as elongation of the helix. In order to project the motion to finite non-zero amplitude, we make use of geometric simulation using the "FRODA" module (25) included in FIRST. FRODA generates new conformations of the protein structure by repeatedly introducing small perturbations of the atomic positions and reimposing the constraints. We use the elastic-network mode eigenvector to bias the perturbations (26); this allows us to project the motion to large amplitudes while maintaining covalent, non-covalent and steric constraints.

The mode can be projected to a C_α RMSD of more than 3 Å from the initial structure (Figure S11a) while maintaining the network of hydrophobic tethers that are present in the original crystal structure. During this projection (Fig. S11b–d) the helix rotates from its initial position diagonal to the sheet to lie parallel to the beta-sheet strands. The projected structures are very similar to conformations from the folding simulation, shown in the main article.

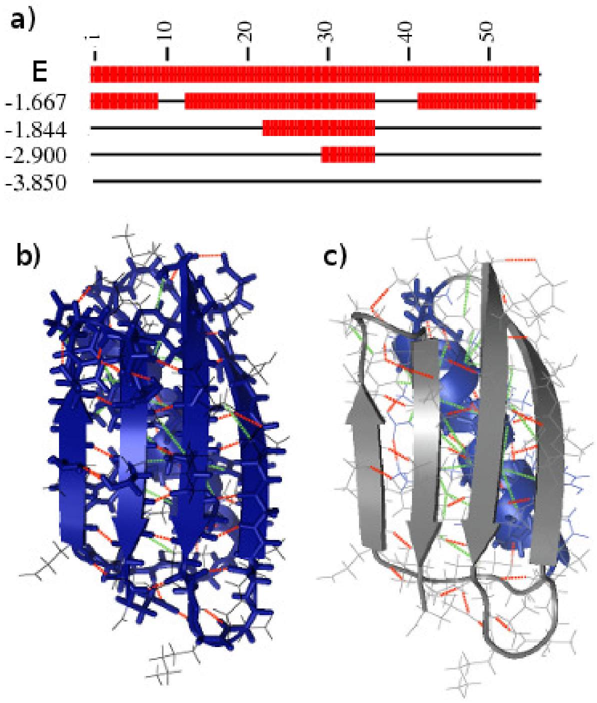


Figure S10: (a) Rigidity dilution of Protein G structure 1PGA as hydrogen bond energy cutoff is lowered. Each line represents the protein backbone; a thin line represents a flexible region while a thick line indicates membership of a large rigid cluster. The beta-sheets lose their rigidity at a cutoff of -1.844 kcal/mol while the helix (residues 22-35) remains rigid to much lower cutoffs. (b) Rigid cluster decomposition of 1PGA at a cutoff of -1.0 kcal/mol. (c) Rigid cluster decomposition at a cutoff of -1.9 kcal/mol. Green and red dashed lines represent hydrophobic tethers and hydrogen bonds.

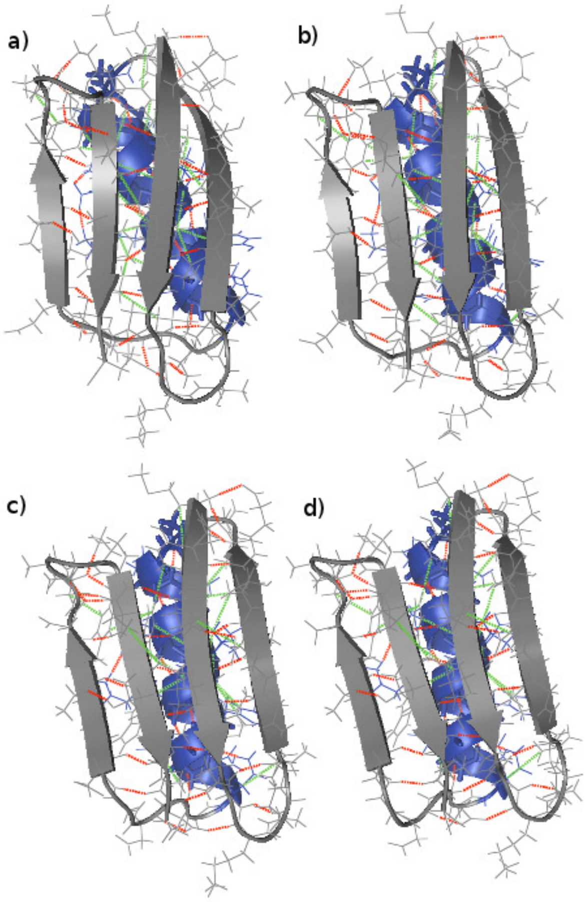


Figure S11: Projection of lowest-frequency non-trivial elastic network mode from initial structure (a) to 3Å RMSD (b–d). Green and red dashed lines represent hydrophobic tethers and hydrogen bonds.

References

[1] Engh, R. A., and R. Huber, 1991. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A* 47:392–400.

[2] Engh, R. A., and R. Huber, 2001. Structure quality and target parameters. In M.G. Rossman, E. Arnold (eds.) International Tables for Crystallography, vol. F: Crystallography of biological macromolecules, 1st edn. Kluwer Academic Publishers for the International Union of Crystallography, Dordrecht Boston London.

[3] Brünger, A., 1992. X-PLOR, Version 3.1: a system for X-ray crystallography and NMR. Yale University Press, New Haven.

[4] Podtelezhnikov, A. A., and D. L. Wild, 2005. Exhaustive Metropolis Monte Carlo sampling and analysis of polyalanine conformations adopted under the influence of hydrogen bonds. *Proteins* 61:94–104.

[5] Ho, B. K., E. A. Coutsias, C. Seok, and K. A. Dill, 2005. The flexibility in the proline ring couples to the protein backbone. *Protein Sci* 14:1011–8.

[6] Srinivasan, R., and G. D. .Rose, 1999. A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* 96:14258–14263.

[7] Brenner, S. E., P. Koehl, and M. Levitt, 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucl. Acids Res.* 28:254–256.

[8] Carreira-Perpinan, M., and G. Hinton, 2005. On contrastive divergence learning. In: Artificial Intelligence and Statistics. The Savannah Hotel, Barbados.

[9] Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan, 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 95–9.

[10] Hopfinger, A. J., 1973. Conformational properties of macromolecules. Academic Press, New York.

[11] Word, J. M., S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, B. K. Presley, J. S. Richardson, and D. C. Richardson, 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* 285:1711–33.

[12] Pappu, R. V., R. Srinivasan, and G. D. Rose, 2000. he Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci U S A* 97:565–70.

[13] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, 2000. The protein data bank. *Nucleic. Acids. Res.* 28:235–42.

[14] Baker, E. N., and R. E. Hubbard, 1984. Hydrogen bonding in globular proteins. *Proc. Biophys. Mol. Biol.* 44:97–179.

[15] Savage, H. J., C. J. Elliott, C. M. Freeman, and J. M. Finney, 1993. Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. *J. Chem. Soc. Faraday Trans.* 89:2609–2617.

[16] Stickle, D. F., L. G. Presta, K. A. Dill, and G. D. Rose, 1992. Hydrogen bonding in globular proteins. *J. Mol. Biol.* 226:1143–59.

[17] McDonald, I. K., and J. M. Thornton, 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238:777–93.

[18] Podtelezhnikov, A. A., Z. Ghahramani, and D. L. Wild, 2007. Learning about Protein Hydrogen Bonding by Minimizing Contrastive Divergence. *Proteins* 66:588–99.

[19] Lee, B., and F. M. Richards, 1971. The interpretation of protein structures estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

[20] Podtelezhnikov, A. A., and D. L. Wild, 2009. Reconstruction and Stability of Secondary Structure Elements in the Context of Protein Structure Prediction. *J. Biophys.* 96:4399–4408.

[21] Gō, N., 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:83–210.

[22] Thorpe, M. F., M. Lei, A. J. Rader, D. J. Jacobs, and L. A. Kuhn, 2001. Flexible and rigid regions in proteins. *J. Mol. Graph. Model* 19:60–69.

[23] Word, J. M., S. Lovell, J. S. Richardson, and D. C. Richardson, 1999. Asparagine and Glutamine: Using Hydrogen Atoms Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* 285:1735–1747.

[24] Suhre, K., and Y.-H. Sanejouand, 2004. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Research (Web Issue)* 32:610–614.

[25] Wells, S. A., S. Menor, B. M. Hespenheide, and M. F. Thorpe, 2005. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* 2:S127–S136.

[26] Jimenez-Roldan, J. E., R. B. Freedman, R. A. Roemer, and S. A. Wells, 2011. Protein flexibility explored with normal modes and geometric simulation. *Phys. Biol.* Accepted.

Chapter 3

Force Field Parameter Inference using Contrastive Divergence

When developing a coarse-grained protein model and force field, such as CRANKITE, it is essential to optimize the parameters in order to produce results that are as accurate and reliable as possible. It is also important to consider the generalizability of the force field, that is, the ability of the force field to correctly model proteins not included in the parameter optimization procedure.

In this paper we optimize the parameters of the CRANKITE force field using the maximum likelihood approach described in the Introduction. We take a set of protein crystal structures from the PDB database as our training set and optimize the force field parameters in order to maximize the likelihood (the room-temperature Boltzmann distribution) of the training set.

Rather than calculating the computationally expensive ensemble averages required for the standard maximum likelihood method, we use a statistical machine learning approach known as contrastive divergence and approximate the direction of the gradient of the logarithm of likelihood. This approach is significantly more efficient and therefore enables a larger and hence more diverse training set to be used. This is in contrast to earlier work, where the training set consisted of only a small number of protein fragments. In that case the optimized force field was not found to be generalizable. However, with the larger training set, we find the optimized force field is generalizable to protein G, a protein not included in the training set.

We compare two different van der Waals potential forms, the standard Lennard Jones 12–6 potential and a computationally cheaper hard cutoff function. Using nested sampling simulations of protein G, we find that the room temperature ensemble of structures using the 12–6 function include the native helix orientation, whereas this is not the case with the hard cutoff van der Waals potential.

We also compare the optimized parameters to those taken directly from ‘standard’ molecular dynamics force fields. Taking the different parameter values, we compare the observed distributions of bond angles, atomic distances, backbone dihedral angles and hydrogen bonding patterns. We also compare heat capacities of polyalanine and the different turn types found when folding a β-hairpin. We demonstrate the importance of optimizing the parameters of the force field rather than taking values found in the literature.

Finally, we discuss the contrastive divergence procedure as applied to force field parameter inference, its behaviour, the assumptions it relies upon and the effect of changing the quality of the training set.

3.1 Contribution

The initial idea was formulated by Wild. Burkoff implemented the initial exploratory tests, which were used by Várnai in order to develop the procedure. Várnai carried out the simulations. Burkoff contributed to the contrastive divergence theory described in the Methods Section and all authors were involved in formulating the discussion in light of the simulation results.

The initial draft of the paper was written by Várnai with Burkoff and Wild contributing to the revision of the paper before submission. Várnai drafted a response to the referees with input from Wild and Burkoff during the peer review process and revised the paper before submission.

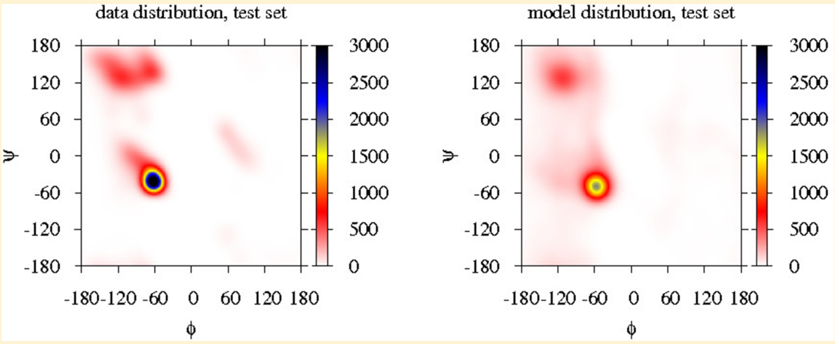
The article was originally published by ACS Publications: Várnai C., N. S. Burkoff and D. L. Wild. Efficient Parameter Estimation of Generalizable Coarse-Grained Protein Force Fields Using Contrastive Divergence: A Maximum Likelihood Approach. *J. Chem. Theory Comput.* (2013) 9(12)5718–5733.

Efficient Parameter Estimation of Generalizable Coarse-Grained Protein Force Fields Using Contrastive Divergence: A Maximum Likelihood Approach

Csilla Várnai, Nikolas S. Burkoff, and David L. Wild*

Systems Biology Centre, University of Warwick, Coventry, United Kingdom

Supporting Information



ABSTRACT: Maximum Likelihood (ML) optimization schemes are widely used for parameter inference. They maximize the likelihood of some experimentally observed data, with respect to the model parameters iteratively, following the gradient of the logarithm of the likelihood. Here, we employ a ML inference scheme to infer a generalizable, physics-based coarse-grained protein model (which includes Gō-like biasing terms to stabilize secondary structure elements in room-temperature simulations), using native conformations of a training set of proteins as the observed data. Contrastive divergence, a novel statistical machine learning technique, is used to efficiently approximate the direction of the gradient ascent, which enables the use of a large training set of proteins. Unlike previous work, the generalizability of the protein model allows the folding of peptides and a protein (protein G) which are not part of the training set. We compare the same force field with different van der Waals (vdW) potential forms: a hard cutoff model, and a Lennard-Jones (LJ) potential with vdW parameters inferred or adopted from the CHARMM or AMBER force fields. Simulations of peptides and protein G show that the LJ model with inferred parameters outperforms the hard cutoff potential, which is consistent with previous observations. Simulations using the LJ potential with inferred vdW parameters also outperforms the protein models with adopted vdW parameter values, demonstrating that model parameters generally cannot be used with force fields with different energy functions. The software is available at https://sites.google.com/site/crankite/.

1. INTRODUCTION

The aim of predicting unknown protein structures from only their primary sequences[1] or to elucidate the folding process or function of proteins with known structures is one of the central aims of computational biology. The increase in the number of protein sequences and structures deposited in the protein databases[2,3] highlights the need for efficient modeling of proteins. Although all-atom molecular force fields have been successfully applied to model fast folding mini-proteins,[4] they are too expensive for modeling larger proteins without the use of specialist hardware. Coarse-grained (CG) protein models, which are simpler than all-atom models, but still capture the physics of interest, have shown an increasing popularity in their use in computer simulations of proteins.[5]

In general, CG force fields are usually classified into two main categories:[5,6] structure-based or native-centric models, such as elastic network[7,8] and Gō models,[9] where only the native interactions are modeled as attractive interactions; and structure-independent force fields[6] that are modeling physicochemical interactions that are often used in simulations of aggregates,[10−12] protein structure prediction,[13] or protein folding studies.[11,14−18] Here, we optimize a Gō-like CG force field, CRANKITE,[19] which was developed to efficiently model peptides and proteins at room temperature by exploiting a fast conformational sampling algorithm,[20] and to stabilize secondary structure elements at room temperature,[21] which would allow it to be used for protein structure prediction[22] using predicted secondary structure and β−β contact maps.[23] It is an extended Gō-type model, where, although some of the secondary structure interactions are constrained using a harmonic bias potential, non-native attractive interactions are also modeled. In

Received: July 18, 2013
Published: November 15, 2013

this paper, the bias potential acting on the backbone conformation of residues with known α-helical and β-strand secondary structure and the β-carbon distances of known β-sheet contacts will be referred to as *secondary structure bias*. Hence, this model allows the exploration of a more realistic folding funnel, compared to the "perfect" funnel of standard Gō models. Thus, CRANKITE represents an intermediate between the two main classes of CG protein models. CRANKITE also uses a full atom representation of the protein backbone, together with explicit side chain β and γ atoms, to include entropic contributions coming from the torsional flexibility of side chains.[24] This is important, because it has been shown that although polyalanine models (including only β atoms) are excellent for modeling secondary structure elements, they form more compact structures than real proteins.[25]

When optimizing force field parameters, protein models should be parametrized to stabilize the native conformation of the protein compared to unfolded and misfolded conformations; that is, the native conformation lies at the global minimum of the free-energy landscape.[26] Traditionally, statistical-knowledge-based potentials have been used to estimate model parameters of the energy function to reproduce certain features of a model dataset,[27] such as dihedral angles and distance distributions, assuming that the selected features are statistically independent and that their distribution in the dataset of native conformations comply with the Boltzmann distribution. This assumption is called the *Boltzmann hypothesis*. Although the Boltzmann hypothesis is supported by numerous empirical studies (see the Discussion section in refs 27 or 28), the assumption of statistical independence is often poor. Moreover, a reference state is usually introduced in the potential of mean force formulation without a rigorous definition, and the decoy sets used to describe the reference state will affect the optimized potential parameters, as demonstrated by Hamelryck et al.,[29−31] who give a rigorous statistical definition of a reference state.

Alternatively, native structure discriminant methods use a set of decoy conformations to optimize the parameter values, such that the folding characteristics of the protein are reproduced, with the lowest energy assigned to the native state, using various optimization techniques.[32−40] However, these methods do not incorporate temperature into the model, and so they do not take into account the thermodynamic stability of proteins, only the relative strength of intermolecular interactions to a set of decoys.

An alternative way of estimating the potential parameters is by using maximum likelihood (ML) methods, which infer the potential parameters by maximizing the likelihood of the experimentally observed (or computationally generated) protein conformations, with respect to the model parameters iteratively (or analytically,[41] for very simple models), following the gradient of the logarithm of likelihood.[12,20,41−44] The model with the parameters giving the highest likelihood would generate a distribution of conformations (*model distribution*) closest to the experimentally observed distribution of conformations (*data distribution*, also referred to as the *target distribution* of the parameter estimation). The free-energy landscape of the inferred model potential is closest to the free energy landscape corresponding to the data distribution, which was demonstrated using a simple model of water,[44] for which the free energies could be calculated analytically. Winther and Krogh,[42] followed by Podtelezhnikov et al.[20,21] used a ML approach to train a protein model (i.e., a model applicable to globular proteins), while Shell et al.[12] used a ML approach to train a protein model specific to a 15-residue polyalanine, a prototype molecule used to model amyloid formation. The relation of this ML approach (also referred to as the relative entropy method[44]) to the force matching method[45] was analyzed by Chaimovich and Shell[46] and Rudzinski and Noid,[47] in the context of fitting CG potentials to all-atom models.

As we show below, the difficulty of the ML approach lies in the calculation of ensemble averages over the model distribution at every iteration. Winther and Krogh[42] conducted extensive simulations using replica exchange molecular dynamics to calculate the ensemble averages, restricting their training set to a small set of short peptides (24 different 11−14-residue-long protein fragments), which resulted in poor transferability to model peptides not in the training set. To efficiently estimate the gradient of the log likelihood, instead of re-evaluating the ensemble averages at each ML iteration, Podtelezhnikov et al.[20,21] used a statistical machine learning technique, known as contrastive divergence (CD),[48] which was developed in the neural network literature to efficiently estimate the parameters of Boltzmann machines.[49,50] This enabled the use of a larger data set of proteins and resulted in a transferable protein model, which was subsequently used in folding simulations of proteins not in the training set. Shell et al.[12] presented another solution to reduce computational costs, using reweighted ensemble averages between successive iterations. To accelerate the convergence of the ML optimization, Hinton[51] suggested an adaptive learning rate with an associated momentum, while Bilionis and Zabaras[52] have proposed an optimization algorithm that makes use of the second derivative of the energy, with respect to the parameters of the energy function.

In our earlier work, we have used the CD algorithm to efficiently estimate potential parameters (hydrogen-bond strength in proteins[20] and the secondary structure bias parameters[21]) of a CG protein model, CRANKITE. The aim of this work is to improve the CRANKITE protein model, as an exemplar for a CG force field, by inferring, or learning, the van der Waals (vdW) parameters of the CG protein model using this statistical machine learning approach. Two potential forms are considered in this paper: a computationally efficient hard cutoff model, employed by the original CRANKITE force field that models short-range repulsion due to the Pauli exclusion between overlapping electron densities, and the Lennard-Jones (LJ) potential form[53] that also models long-range attraction due to fluctuating charge densities of induced dipoles. Following the explanation of the method, the parameter inference and the effect of the simulation parameters on the inference are discussed. Subsequently, the improvement of the force field is investigated by a comparison of the performance of the hard cutoff and LJ type potential forms through the investigation of structural and thermodynamic properties, calculated from Monte Carlo (MC) and folding simulations of 16-residue peptides and protein G (Protein Databank (PDB) code: 1PGA). Transferability between different protein models is tested by comparisons of LJ type potentials with learnt vdW parameters ($\mathrm{LJ}_{\mathrm{learnt}}$) and parameters adopted from the widely used AMBER[54] and CHARMM[55] all-atom force fields ($\mathrm{LJ}_{\mathrm{AMBER}}$ and $\mathrm{LJ}_{\mathrm{CHARMM}}$, respectively). The assumptions of the method are also discussed.

2. METHODS

2.1. Maximum Likelihood Inference for Parameter Estimation of Generalizable Protein Models. We assume that we have n_0 independent observations of the conformation Ω_0 of a protein with amino acid sequence S_0, $\{\Omega_0|S_0\} = \{\Omega_0^j|S_0: j = 1, ..., n_0\}$, distributed according to the Boltzmann distribution at inverse temperature β (e.g., the outcomes of an experiment or a computer simulation). The interaction parameters, θ, of a protein model, such as force constants, distance cutoffs, dielectric permittivity or atomic charges, specific to the protein with amino acid sequence S_0, can be estimated by maximizing the likelihood, $L = P(\theta|\{\Omega_0\},S_0)$, by a gradient ascent using an iterative scheme. At iteration k+1,

$$\theta^{k+1} = \theta^k + \eta\nabla_\theta \ln L \tag{1}$$

where η is the learning rate, and ∇_θ ln L is the gradient of the logarithm of likelihood, with respect to parameter θ. Assuming that the observations $\{\Omega_0\}$ are independent and come from the Boltzmann distribution at inverse temperature β for a given parameter set θ,

$$P(\{\Omega_0\}|\theta, S_0) = \prod_{j=1}^{n_0} \frac{\exp(-\beta E(\Omega_0^j|\theta, S_0))}{\int \exp(-\beta E(\Omega|\theta, S_0))\mathrm{d}\Omega} \tag{2}$$

Using Bayes' equality with a uniform prior $P(\theta|S_0)$, the gradient of the likelihood, with respect to the model parameters, can be written as

$$\nabla_\theta \ln L = -n_0\beta\left(\frac{1}{n_0}\sum_{j=1}^{n_0} \nabla_\theta E(\Omega_0^j|\theta, S_0) - \langle\nabla_\theta E(\Omega|\theta, S_0)\rangle_{\theta,S_0}\right) \tag{3}$$

where $\langle A(\Omega)\rangle_{\theta,S0} = \int A(\Omega)P(\Omega|\theta,S_0)\mathrm{d}\Omega$ is the ensemble average of $A(\Omega)$ in the model distribution. The first term in the parentheses of eq 3 is an average over the data, approximating an ensemble average over the data distribution. Maximizing the likelihood is equivalent to minimizing the Kullback–Leibler divergence (or relative entropy) of the data distribution and the model distribution:

$$\begin{aligned}\mathbf{KL}(P(\Omega_0|S_0)||P(\Omega_0|\theta, S_0)) &= \sum_{j=1}^{n_0} P(\Omega_0^j|S_0)\ln\left(\frac{P(\Omega_0^j|S_0)}{P(\Omega_0^j|\theta, S_0)}\right)\\ &= -H(P(\Omega_0)) - \frac{1}{n_0}\sum_{j=1}^{n_0} P(\Omega_0^j|S_0)\\ &\quad\times \ln P(\Omega_0^j|\theta, S_0)\end{aligned} \tag{4}$$

since the entropy of the data distribution, $H(P(\Omega_0)) = -\sum_{j=1}^{n_0}P(\Omega_0^j|S_0)\ln P(\Omega_0^j|S_0)$, does not depend on the parameters θ, and the observations are drawn from the data distribution.

Such a protein model will be specific to the protein with sequence S_0 it was trained on, and is unlikely to be transferable to proteins with arbitrary amino acid sequences. A generalizable protein model, that is, one that is transferable to proteins not in the dataset, must be trained on a set of proteins that are representative of all the proteins we aim to model, and which are independent of each other. Hence, let us take observations of the conformations of N proteins with amino acid sequences $\{S_0\} = \{S_0^i: i = 1, ..., N\}$. Let us allow that, for some proteins with sequence S_0^i, more than one independent observation of the conformation is available, $\{\Omega_0^i\} = \{\Omega_0^{ij}|S_0^i: j = 1, ..., n_i\}$, and that all observations come from the Boltzmann distribution corresponding to the same inverse temperature β. The parameters of the generalizable protein model (we use the same parameter set θ to describe all proteins) maximize the likelihood of the parameters, given the observed conformations. The probability of finding the dataset, given the sequences and the parameters, is

$$P(\{\{\Omega_0^i\}: i = 1, ..., N\}|\theta, \{S_0\}) = \prod_{i=1}^{N} P(\{\Omega_0^i\}|\theta, S_0^i) \tag{5}$$

as a conformation is only dependent on its own protein sequence and the general θ parameters. Following a similar derivation to that for eq 3, the gradient of the logarithm of likelihood, with respect to the model parameters θ, can be written as

$$\nabla_\theta \ln L = -\beta\sum_{i=1}^{N}\left[n_i\left(\frac{1}{n_i}\sum_{j=1}^{n_i}(\nabla_\theta E(\Omega_0^{ij}|\theta, S_0^i)) - \langle\nabla_\theta E(\Omega|\theta, S_0^i)\rangle_{\theta,S_0^i}\right)\right] \tag{6}$$

This is equivalent to minimizing the average of KL divergences between the data and model distributions for all sequences,

$$\sum_{i=1}^{N} \mathbf{KL}(P(\Omega_0^i|S_0^i)||P(\Omega_0^i|\theta, S_0^i)) \tag{7}$$

Note that neither the length of the proteins, nor other properties of the protein sequences explicitly affect the parameter estimation; the direction of the gradient ascent is given by the unbiased average of the KL divergences of the model and data distributions for all sequences S_0^i. Also note that if there is only one observation available for any protein sequence, the first term of the inner sum of eq 6, the average over the data distribution for S_0^i, is approximated by one data point. Even in this case, the ML estimate is still correct, as long as all protein conformations are described by the Boltzmann distribution at the same inverse temperature β, and they are representative of the proteins we aim to model.

2.2. Contrastive Divergence. In contrastive divergence,[48] to avoid the cumbersome calculation of the ensemble average in the model distribution at every step of the ML iteration (eq 6), the Kullback–Leibler divergence of the data distribution and a perturbed data distribution is minimized, instead of the KL divergence of the model and data distributions. Samples from the perturbed distribution are generated by performing K MC steps starting from the observed conformations representing the data distribution, using the model parameters θ^k at iteration k. For a protein with amino acid sequence S_0, we use $P^0(\Omega|S_0) = P(\Omega_0|S_0)$ to denote the data distribution, $P_\theta^\infty(\Omega|S_0) = P(\Omega|\theta,S_0)$ to denote the equilibrium distribution of the model with parameters θ, and $P_\theta^K(\Omega|S_0)$ to denote the perturbed data distribution, which is generated by performing K MC steps starting from the data distribution using the model parameters θ at every iteration. The direction of gradient ascent is given by

$$\frac{\partial[\mathbf{KL}(P^0(\Omega|S_0)||P_\theta^\infty(\Omega|S_0)) - \mathbf{KL}(P_\theta^K(\Omega|S_0)||P_\theta^\infty(\Omega|S_0))]}{\partial\theta} = \left\langle \frac{\partial E(\Omega|\theta, S_0)}{\partial\theta} \right\rangle_K - \left\langle \frac{\partial E(\Omega|\theta, S_0)}{\partial\theta} \right\rangle_0 - \frac{\partial P_\theta^K(\Omega|S_0)}{\partial\theta} \frac{\partial \mathbf{KL}(P^0(\Omega|S_0)||P_\theta^\infty(\Omega|S_0))}{\partial P_\theta^K(\Omega|S_0)} \tag{8}$$

where $\langle A(\Omega|S_0)\rangle_0 = (1/n_0) \times \sum_{j=1}^{n_0} A(\Omega_0^j|S_0)$ is the ensemble average in the data distribution, and $\langle A(\Omega|\theta, S_0)\rangle_K = (1/n_0) \times \sum_{j=1}^{n_0} A(\Omega_K^j|\theta, S_0)$ is the corresponding average in the perturbed data distribution, with Ω_K^j being a conformation in the perturbed data distribution. In the original work by Hinton,[48] simulation results of restricted Boltzmann machines with a small number of visible and hidden units demonstrate that the third term may be safely ignored, and so the CD parameter estimation algorithm becomes

$$\theta^{k+1} = \theta^k + \eta\beta\left[\left\langle \frac{\partial E(\Omega|\theta^k, S_0)}{\partial\theta} \right\rangle_K - \left\langle \frac{\partial E(\Omega|\theta^k, S_0)}{\partial\theta} \right\rangle_0\right] \tag{9}$$

For the problem at hand, we additionally provide the following argument. As $K \to \infty$, eq 1 is recovered. However, even for a small number of steps, unless the model distribution reproduces the data distribution, $P_\theta^K(\Omega|S_0)$ drifts away from the data distribution, toward the model distribution $\mathbf{KL}(P^0(\Omega|S_0) \| P_\theta^\infty(\Omega|S_0)) > \mathbf{KL}(P_\theta^K(\Omega|S_0) \| P_\theta^\infty(\Omega|S_0))$, and the drift in the energy gradient observed during the MC simulation can be used as the estimate of $\nabla_\theta \ln L$. Changing the parameters according to eq 9 reduces the tendency of the model distribution to drift away from the data distribution. To support this argument for the convergence of the algorithm using the approximate gradient, we calculated the distribution of the approximate $\nabla_\theta \ln L$ for different model parameter values, and plotted the distributions at the initial and converged values of one of the model parameters (Figure S1 in the Supporting Information). The expected value of the distribution at the initial parameter values is nonzero (and has the correct sign), while at the converged parameter values, it is zero.

When the observed conformations belong to proteins with different amino acid sequences (i.e., when inferring a generalizable protein model with $n_i = 1$ for all sequences S_0^i), the ML algorithm takes the form

$$\theta^{k+1} = \theta^k + \eta\beta \sum_{i=1}^{N} (\langle \nabla_\theta E(\Omega^i|\theta, S_0^i)\rangle_K - \nabla_\theta E(\Omega^i|\theta, S_0^i)) \tag{10}$$

when using the CD estimation of the KL divergences for all proteins with amino acid sequence S_0^i. This equation is used throughout this work in the ML inference of the protein model parameters. As a constant during the ML inference, β can be incorporated into the learning rate.

2.3. The Protein Model. We use a protein model with an all-atom backbone and coarse-grained side chains represented up to the gamma atoms, as described by Podtelezhnikov et al.[20] and Burkoff et al.[22] Bond lengths and bond angles are rigid, with values taken from Srinivasan et al.[56] and Burkoff et al.,[22] except for the C_α valence angle τ (the angle determined by the amide N, C_α and carbonyl C atoms of a residue), which is allowed to change. Peptide bond geometries are kept fixed, resulting in fixed C_α–C_α distances. The conformational flexibility of the backbone comes from free rotation around the φ and ψ dihedral angles and the C_α valence angle. The side-chain (N–C_α–C_β–C_γ) dihedral angles can take values of ±60° or 180°. During MC simulations, the move set consists of crankshaft rotations around any axes connecting up to 4 C_α atoms, and rotations at the termini around any axis passing through the C_α atom, as implemented in the CRANKITE software.[19,57] At every fourth MC step, the side-chain dihedral angles were reassigned by drawing from the frequency distribution of side chain dihedral angles in the dataset.

The energy function of the protein model depends on the conformation Ω containing all coordinates of its N residues, and the parameter set θ. It consists of six terms,[22]

$$\begin{aligned} E(\Omega, \theta) &= E^{\mathrm{B}} + E^{\mathrm{vdW}} + E^{\mathrm{HB}} + E^{\mathrm{hyd}} + E^{\mathrm{SC}} + E^{\mathrm{P}} \\ &= (\sum_{l=1}^{N} k_\tau(\tau_l - \tau_0)^2) + E^{\mathrm{vdW}} \\ &\quad + \left[\sum_{l=1}^{N} \sum_{m=1,|l-m|>2}^{i} H(n_{l\to m}^{\mathrm{HB}} + n_{m\to l}^{\mathrm{HB}})\right] \\ &\quad + (\sum_{l=1}^{N}\sum_{m=1}^{l} M_{\mathrm{lm}} k_h f_{\mathrm{cut}}^{\mathrm{hyd}}) \\ &\quad + \left[-\sum_{l=1}^{N-1} \eta_{ss_i} \cos((\gamma_{l,l+1} - \gamma_{0,ss_l})) \right.\\ &\quad \left. + \sum_{l=1}^{N}\sum_{m=1}^{l} \kappa_\beta C_{lm}(r_{lm} - r_{0,\beta})^2\right] + (\sum_{l=1}^{N} k_{\mathrm{P}}(\phi_l - \phi_0)^2) \end{aligned} \tag{11}$$

E^{B} is the backbone stress term due to deviations of the C_α valence angle τ_i of residue i from the equilibrium value τ_0 = 69°,[58] and k_τ is the force constant of the quadratic potential. E^{vdW} is the van der Waals interaction term described below, employed to prevent atomic clashes, and to model long-range weak attractive interactions. E^{HB} is the hydrogen bonding term with hydrogen-bond strength H. $n_{l\to m}^{\mathrm{HB}}$ is a number between 0 and 1 representing the strength of hydrogen bonding between the amide H atom of residue l (H_l) and the carbonyl O atom of residue m (O_m), determined using a distance cutoff δ and two angle cutoffs (Θ_{COH} and Ψ_{OHN}). (For the exact function form, see the Supporting Information.) E^{hyd} is a hydrophobic interaction term with interaction strength k_h, a hydrophobic match factor M_{lm}, and the cutoff function $f_{\mathrm{cut}}^{\mathrm{hyd}}$. The hydrophobic match takes a value of 2 if both amino acids are hydrophobic, 1 if one is hydrophobic and the other one is amphipathic, and 0 otherwise. The cutoff function changes linearly from 1 to 0 as the distance of the C_β atoms of residues l and m goes from the sum of vdW radii (from the hard cutoff model) across 2.8 Å. E^{SC} is the side-chain–side-chain interaction term representing a secondary structure bias on the dihedral angles of the residues as well as β-sheet contacts. The $\gamma_{l,l+1}$ dihedral angle, N_l–$C_{\alpha,l}$–$C_{\alpha,l+1}$–C_{l+1}, is restrained to an equilibrium value $\gamma_{0,\mathrm{ss}}$ typical for the corresponding secondary structure element ss ($\gamma_{0,\alpha}$ = 82° for α-helical conformation, and $\gamma_{0,\beta}$ = 180° for β strand conformation) using the force constant η_{ss} (η_α for residues in an α-helical, and η_β for residues in β-strand conformation, and 0 otherwise, defined by a predetermined secondary structure). The $C_{\beta,l}$–$C_{\beta,m}$ distances of residues l and m (r_{lm}) that are in β-sheet contact, defined by a predetermined binary contact map C_{lm} are restrained by a quadratic potential to an equilibrium value $r_{0,\beta}$ using a force constant κ_β. E^{P} is a proline term, specific due to deviations of the C_{l-1}−N_l−$C_{\alpha,l}$−C_l dihedral angle, ϕ_l, of the proline residue l from the equilibrium value of $\phi_0 = -60°$,[59] and $k_P = 30RT$ is the force constant of the quadratic potential.[20]

In this paper, we consider the following forms of the vdW interactions acting between atoms:

- A hard cutoff potential, often used by CG models, because of its simplicity and computational efficiency,[56,57,60] with a distance-dependent excess energy for clashing atoms:[22,57]

$$E_{ij}^{\mathrm{vdW}}(r_{ij}) = \begin{cases} 10\left(\dfrac{0.95(R_i + R_j)^2}{r_{ij}^2}\right)RT & r_{ij} \le 0.95(R_i + R_j) \\ 0 & r_{ij} > 0.95(R_i + R_j) \end{cases} \qquad (12)$$

where R_i and R_j are the vdW radii of atoms i and j taken from the original CRANKITE model,[20] and r_{ij} is their distance.

- A Lennard-Jones potential form, also used in more sophisticated CG models.[32,61,62] Here, the vdW energy is

$$E_{ij}^{\mathrm{vdW}}(r_{ij}) = \begin{cases} \max\left\{\varepsilon_{ij}\left[\left(\dfrac{R_{\min,ij}}{\sqrt{0.4}(R_i + R_j)}\right)^{12} - 2\left(\dfrac{R_{\min,ij}}{\sqrt{0.4}(R_i + R_j)}\right)^{6}\right]; 30RT\right\} & \text{for } r_{ij} \le \sqrt{0.4}(R_i + R_j) \\ \varepsilon_{ij}\left[\left(\dfrac{R_{\min,ij}}{r_{ij}}\right)^{12} - 2\left(\dfrac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] & \text{for } r_{ij} > \sqrt{0.4}(R_i + R_j) \end{cases} \qquad (13)$$

between atoms i and j at a distance r_{ij}, where ε_{ij} is the vdW energy contribution at the minimum energy separation, $R_{\min,ij}$ (see Figure S2 in the Supporting Information). The energies are shifted to obtain zero vdW energy at the cutoff, $2R_{\min,ij}$. For simplicity, the ε parameters of the LJ model are kept the same for all atom types.

More-sophisticated approximations of the vdW potential (for example, the Buckingham potential[63] or many-body Axilrod−Teller−Muto contributions[64]) would be computationally too expensive to include in our CG simulations, where the aim is to develop the simplest protein model that captures the physics of the systems of interest.

2.4. The Optimization Procedure.

In this work, the following parameters of the energy function (eq 11) were inferred for all models considered: the backbone stress force constant (k_τ), the hydrogen-bond strength (H), the hydrogen-bond distance cutoff (δ) and angle cutoffs (Θ_{COH} and Ψ_{OHN}), the hydrophobic interaction strength (k_h), the secondary structure biasing dihedral angle force constants (η_α and η_β), and the C_β−C_β contact equilibrium distance ($r_{0,\beta}$) and force constant (κ_β). For the hard cutoff model, no further parameters were inferred. For the LJ model (eq 13), a mutual vdW energy contribution ε_i parameter for all atom types and the minimum energy separation parameters $R_{\min,i}$ for every atom type (CA, CB, C, N, O and S) were also inferred ($\mathrm{LJ_{learnt}}$ model), or adapted from the CHARMM and AMBER force fields ($\mathrm{LJ_{CHARMM}}$ and $\mathrm{LJ_{AMBER}}$; see Table 1). Note that, in the $\mathrm{LJ_{learnt}}$ model, the CRANKITE atom types have the same ε_i parameter, while in the $\mathrm{LJ_{CHARMM}}$ and $\mathrm{LJ_{AMBER}}$ models they have individual ones.

During the ML inference, the parameters were inferred in two stages, following a multigrid approach.[65] The potential parameters that govern the local and global configurations are separated in the inference, starting with local parameters, and then moving to more global parameters. The local parameters were chosen as those affecting the local configuration of atoms and short atomic distances near atomic clashes, namely, the hydrogen bonding (H, δ, Θ_{COH}, and Ψ_{OHN}), C_α valence angle stress (k_τ), and vdW potential parameters ($R_{\min,i}$ and ε, only for the LJ model), and were inferred together in the first stage. The other parameters acting over larger distances, namely, the secondary structure bias (η_α, η_β, κ_β, and $r_{0,\beta}$), and hydrophobicity (k_h), were learnt subsequently, with the former ones being fixed. Note that the LJ potential also acts at long distances, and hence, the length scale separation is not perfect. In fact, it is an effective way to decouple the effects of potentially competing long-range parameters, such as the hydrophobic interaction potential or the C_β−C_β contact potential, from the short-range part of the LJ parameters, which cannot be decoupled from the long-range part of the LJ parameters.

As the data set of known protein structures representing thermodynamic equilibrium, we use a subset of the protein structures in the ASTRAL 1.75 database.[66] To avoid proteins with high sequence similarity, proteins with less than 40% sequence identity were included. The ASTRAL 1.75 database contains three-dimensional (3D) structures of protein domains, classified into folding classes. For each structure, a Summary PDB ASTRAL Check Index (SPACI)[67] score is assigned, indicating the reliability of crystallographically determined structures. All PDB structures from the α, β, α+β, and α/β classes of the ASTRAL 1.75 database with SPACI scores above 0.8 were included in the dataset, excluding the ones with missing residues, disulfide bonds, or unusual residues.

Following the inference, the hydrophobic interaction strength k_h needed modification. k_h was increased by 0.1 RT, which was necessary for the protein folding simulations to stabilize the conformation with the hydrophobic residues in the interior of the protein. Although the hydrophobic interaction strength was sufficient to preserve the folded structure of the proteins in the database, it was not sufficiently strong for folding proteins from an unfolded state. A possible reason for the learnt value of k_h being too small could be that the ASTRAL

Table 1. The CHARMM and AMBER Atom Types Whose LJ Parameters Were Adopted for the CRANKITE Atom Types in the $\mathrm{LJ_{CHARMM}}$ and $\mathrm{LJ_{AMBER}}$ Models

CRANKITE	CA	CB	C	N	O	S
CHARMM	CT1	CT2	C	NH2	O	S
AMBER	CT	CT	C	N	O	S

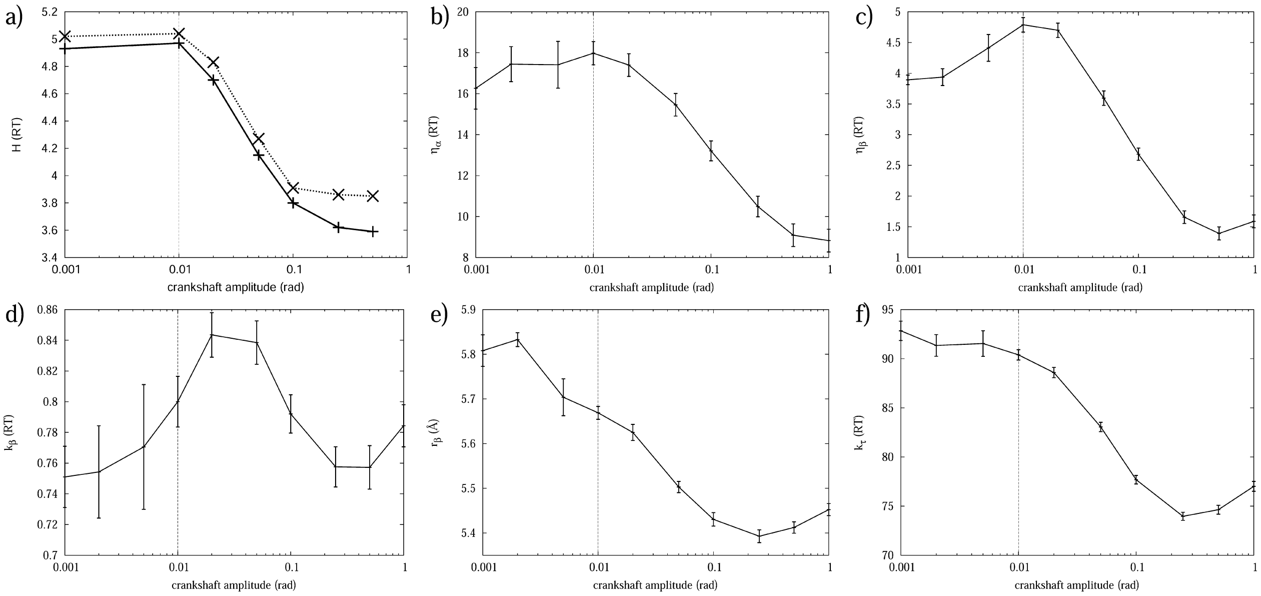


Figure 1. Dependence of the converged potential parameter values, as a function of the Monte Carlo (MC) step size, inferred using the ASTRAL PDB structures after removing overlapping atoms (solid lines), thus using a dataset that better represents the Boltzmann distribution. The plots correspond to (a) hydrogen-bond strength (H), (b) α-helix backbone dihedral angle bias potential strength (η_α), (c) β-strand backbone dihedral angle bias potential strength (η_β), (d) β–β contact bias potential strength (κ_β), (e) β–β contact equilibrium distance ($r_{0,\beta}$), and (f) C_α valence angle stress potential strength (k_τ). For the hydrogen-bond strength plot (panel a) only, parameter values inferred using the ASTRAL PDB structures without removing overlapping atoms are also shown (represented by a dotted line). Vertical dashed lines mark a crankshaft MC step size of 0.01. The error bars correspond to one standard deviation of the distribution of the converged parameter value.

1.75 database used contains individual domains of multidomain proteins, thus including numerous hydrophobic residues on the surfaces of proteins in the dataset, although these would be in the interior of the native multidomain proteins. Moreover, increasing the hydrophobic interaction strength in effect incorporates a penalty term for hydrophobic–hydrophilic interactions of hydrophobic side chains with water molecules. All other potential parameters were used unmodified.

The convergence was monitored by calculating the mean and the standard deviation of parameter values for consecutive 1000-step intervals. When the mean changed by less than the standard deviation and it fluctuated over three consecutive steps, convergence was achieved. The simulations were further run for another 5000 steps, and from these steps, the mean and standard deviation of the distributions of the parameter values were calculated.

2.5. Simulation Parameters.

For the parameter estimation, structures in the protein database were mapped onto the protein model. In the mapping process, in which constraints of the CG model are enforced, a few atomic clashes are introduced. In order to eliminate high-energy configurations due to clashing atoms, the following modifications were made to the PDB library. The C_β–C_γ distances of amino acids with long and flexible side chains (lysine, methionine, glutamine, and arginine) were set to their real C_β–C_γ bond lengths: 1.52 Å for lysine, methionine, and arginine, and 1.53 Å for glutamine. Furthermore, any γ atoms that caused atomic clashes (for instance, due to nonstandard side-chain dihedral angles), 765 atoms in total, were removed from the PDB structures used. Subsequently, PDB structures whose backbone atoms were involved in further atomic clashes after the mapping onto the protein model, 6 proteins in total, were also removed from the library. The list of the proteins used with their SPACI scores, ASTRAL class information, and the α-carbon root-mean-square distance (RMSD) of the mapped and the original structures are included in Table S1 of the Supporting Information. The maximum C_α RMSD between a mapped and an original structure was 0.045 Å, while the mean C_α RMSD between the mapped and the original structures was 0.025 Å.

In the CD learning simulations, we use 4096 MC moves per CD learning iteration, and a temperature of 298 K was used in calculating the Metropolis–Hastings acceptance criterion. The learning rate of the CD learning simulations for each parameter was determined by a trial-and-error method and set to be sufficiently large to speed up the convergence, but small enough to avoid instabilities in the convergence. The effect of the maximum amplitude of the crankshaft rotations during the CD learning was also investigated (see the Results section).

To validate the model parameters against the data, the model distributions of some geometric observables using the optimized parameters were compared to the data distribution of the training set. The model distributions were generated by 10^6 step MC simulations using the protein models with optimized parameters, starting from the training set, or from an independent PDB set consisting of structures of the ASTRAL 1.75 database with SPACI scores between 0.7 and 0.8.

The inferred vdW potentials were further tested using 16-residue peptides and a 56-residue protein, Protein G (1PGA). First, a 10^8 step MC simulation was performed on a 16-residue polyalanine peptide, using only the stress, vdW, and hydrogen-bond contributions of the energy function, to determine the accessible areas on the Ramachandran map, and the stable secondary structure forms without using any secondary structure bias. Subsequently, nested sampling (NS)[22,68] simulations of β-hairpin folding were performed on a 16-residue polyalanine and its glycine mutants, introducing a β-

Table 2. Inferred Potential Parameters Using Contrastive Divergence, for the Protein Models Using the Hard Cutoff and the Lennard-Jones (LJ)-Type van der Waals (vdW) Potentials[a]

vdW and Backbone Stress Potential Parameters

vdW potential	$R_{\min}^{CA}$	$R_{\min}^{CB}$	$R_{\min}^{C}$	$R_{\min}^{N}$	$R_{\min}^{O}$	$R_{\min}^{S}$	$\varepsilon(RT)$	$k_\tau(RT)$
hard cutoff	1.57	1.57	1.42	1.29	1.29	2.00		90
LJ_{learnt}	2.43	1.97	1.82	1.74	1.98	3.10	0.018	98
LJ_{CHARMM}	2.275	2.175	2.00	1.85	1.70	2.00	[b]	103
LJ_{AMBER}	1.908	1.908	1.908	1.824	1.6612	2.00	[c]	114

Hydrogen-Bond Potential Parameters

vdW potential	H (RT)	δ (Å)	cos Θ_{COH}	cos ψ_{OHN}
hard cutoff	4.95	2.01	0.770	0.930
LJ_{learnt}	4.98	2.01	0.772	0.928
LJ_{CHARMM}	4.80	2.01	0.772	0.925
LJ_{AMBER}	4.91	2.01	0.771	0.921

Secondary Structure Bias Potential Parameters

vdW potential	η_β (RT)	η_α (RT)	K_β (RT/Å^2)	R_β (Å)
hard cutoff	4.5	18.0	0.80	5.65
LJ_{learnt}	3.7	15.3	0.85	5.39
LJ_{CHARMM}	4.5	18.6	1.00	5.62
LJ_{AMBER}	2.6	19.7	1.18	5.15

Hydrophobic Interaction Potential Parameters

vdW potential	k_h (RT)
hard cutoff	0.030
LJ_{learnt}	0.022
LJ_{CHARMM}	0.051
LJ_{AMBER}	0.057

[a] The vdW potential parameters of the hard cutoff model were taken from ref 20, while those of the LJ_{CHARMM} and LJ_{AMBER} models were taken from the CHARMM[55] and AMBER[54] force fields, respectively. [b] ε/RT values from the CHARMM force field (0.0338, 0.0929, 0.186, 0.338, 0.203, and 0.760 for the CA, CB, C, N, O, and S atom types respectively). [c] ε/RT values from the AMBER force field (0.185, 0.185, 0.145, 0.287, 0.355, and 0.422 for the CA, CB, C, N, O, and S atom types respectively). The potential parameters are described in section 2.3; wherever a unit of length is not indicated, the unit of length is Å.

hairpin secondary structure bias, to examine the behavior of the unbiased loop. In the mutants, a glycine residue was introduced at amino acid positions 8, 9, or 10, corresponding to the i+1, i+2, and i+3 positions in the turn. Nested sampling is a Bayesian sampling technique,[68] which has been shown to be superior to parallel tempering with regard to finding the native basin of Protein G using the CRANKITE protein model in our previous work.[22] Further NS simulations were performed on the 16-residue polyalanine peptide using α-helix and β-hairpin secondary structure bias, respectively, to determine melting heat capacity curves of the secondary structure. The NS simulations were performed until the partition function converged to $T = -100$ °C, which implies that the thermodynamically accessible states have been sampled for all temperatures above T, and hence, the heat capacity values have been converged for any temperature above −100 °C. In the NS simulations of the 16-residue peptides, 10 000 active points were used, and 10 000 MC steps were used to generate new points in the NS iterations. In the NS simulations of Protein G, 20 000 MC steps and 20 000 active points were used, and the partition function was converged down to 25 °C.

3. RESULTS

3.1. Effect of the Simulation Parameters on the Inference.

In a contrastive divergence iteration, a short MC simulation is performed to estimate the gradient of the energy, with respect to the simulation parameters. The number of MC steps, K, during each CD iteration affects the quality of the gradient estimation, that is, the smaller the K value, the more stochastic the gradient estimate becomes; however, K does not affect the overall maximum likelihood.[48] A more-stochastic estimate of the gradient slows the convergence of the CD simulations; however, it will not prevent convergence. Following an argument by Hinton,[48] even for K as small as 1, on average, over the training set, the perturbed data distributions are closer than the data distribution to the equilibrium distribution of the current model parameters (unless the data and model distributions are equal), even if individual MC simulations might result in an opposing gradient at any iteration. Throughout this work, we use K = 4096, which was found to be effective for the parameter inference.

During the MC evolution of each CD iteration, the maximum allowed amplitude of the crankshaft rotations affects the local exploration, thus influencing the converged potential parameter values (Figure 1), and this can cause significant variations in the converged parameter values. Our aim is to infer a protein model that can be used in protein folding simulations; hence, the exploration must be local for the quadratic functions to describe the local basin, but it should also be able to describe the energy surface nonlocally, and not only the energy restrained to the crystal structure. In this work, we approximate many terms of the energy function using quadratic functions. On rugged energy landscapes where this harmonic approximation of the curvature of the landscape is a very crude approximation, larger MC moves facilitate the crossing (effectively tunnelling) of energy barriers that smaller MC moves could not climb over, and this makes the potential energy surface appear to be different, often flatter (e.g., increasing the MC step size from 0.01 to 0.2 in Figures 1a, 1b, 1c, and 1f). Since the parameter estimates do not change by more than 5% for amplitudes of 0.001−0.01 radians for H and k_τ and 0.002−0.02 radians for η_α, in the following, we chose to use a maximum crankshaft rotation of 0.01 radians in the CD estimations of the parameters for all models, and we will be comparing results using this maximum rotation amplitude. For the other parameters, we accept that the harmonic approximation is probably far from perfect.

We also note that, although the convergence of parameters for the individual maximum amplitude sizes is not prohibited, the speed of convergence also depends on the MC step size. Decreasing the MC step size increases the acceptance rate, although from an MC step size of 0.02, the acceptance rate is over 80% (see Figure S3 in the Supporting Information), and it does not give much advantage in the exploration of the energy surface during the short MC simulations used to estimate the gradient in the CD iterations. On the other hand, when the allowed MC step size is set to be small (for a given number of MC steps), the exploration of the energy surface becomes poorer, and the poorer gradient estimate slows the convergence of the CD simulations.

3.2. Estimation of the Protein Model Parameters. When inferring several potential parameters together, learning correlated potential parameters is crucial for the convergence of the ML estimation. This can be done by considering the functional form of the energy function. When using the LJ-type potential that is designed to have a nonspecific long-range attractive energy contribution, we find problems with the convergence of the CD learning of the parameters. The reason for this is that the attractive interactions of the LJ potential compete with the short-range attractive interactions. For example, the vdW interaction between a N atom and an O atom of a hydrogen bond would compete with the hydrogen-bond interaction between them, both trying to describe an attractive interaction between the two atoms at the same time. Similarly, distances that occur frequently in secondary structure elements (and are therefore enforced by the secondary structure bias interactions), e.g., the C_β−C_β distance of interacting amino acid residues in a β-sheet, would introduce an artificial bias to the LJ potential parameters. To avoid these problems, we only evaluate the hard cutoff part of the LJ potential between atoms of amino acid residues that are connected via a hydrogen bond, or whose neighbors are connected via a hydrogen bond. This way, only nonspecific nonbonded interactions are taken into account in the parameter estimation of the LJ potential, and the correlation of the potential parameters are suppressed for the inference. Other ways to address this problem include fixing a parameter value, or the ratio of the competing parameters together (e.g., merging the hydrogen bond with a hydrophobicity into one function). However, introducing such constraints on the potential parameters could introduce an artificial bias on the parameter values.

The inferred values of the vdW potentials, hydrogen bond, secondary structure bias, and hydrophobicity potential parameters are summarized in Table 2, together with corresponding values taken from the CHARMM and AMBER force fields. While there is no noticeable difference between the hydrogen-bond potential parameters for the two vdW models, the force constant k_τ of the backbone stress interaction is higher for the protein model using the LJ potential than for the one using the hard cutoff potential. This indicates that when using the LJ functional form, as opposed to the hard cutoff functional form, to represent the atoms, a larger conformation space might be available by applying the vdW potential, and a higher backbone stress force constant compensates for this, to obtain the equilibrium distribution of C_α valence angles in the dataset. This is supported by the comparison of vdW interaction functions between various atom types using the model parameters. Also, the β-strand backbone bias potential parameter, η_β, is noticeably higher for the protein model using the hard cutoff vdW potential, which shows that the LJ models favor the extended conformation more than the hard sphere model. On the other hand, the β−β contact potential is slightly stronger in the protein model using the LJ-type vdW potential, with shorter equilibrium distance, $r_{0,\beta}$, for the interacting C_β atoms, and a slightly higher force constant. The α-helices might also be slightly more stable without a bias potential, suggested by the lower α-helix backbone bias force constant, η_α.

During the ML inference, the KL divergence of the model and data distributions is minimized. However, for an unrealistic energy function, the model distribution might still be far from the data distribution. To validate our protein models for describing the training set of proteins, we calculate various structural observables in the model and data distributions, such as the backbone dihedral angles (see Figures S4 (left) and S5 in the Supporting Information), the α-carbon valence angle (see Figure S7 (left) in the Supporting Information) and the distribution of the distance between β-carbon atoms of interacting amino acid residues in β-sheets (see Figure S7 (left) in the Supporting Information). Although the above distributions are 1-dimensional (1D) or two-dimensional (2D) marginalizations of the joint distributions, they would provide a good indication if the model distribution were different from the data distribution. In our current work, all model distributions of the α carbon valence angle are identical to the data distribution. The model distribution of the β-carbon atoms of interacting amino acid residues in β-sheets in the $\mathrm{LJ}_{\mathrm{AMBER}}$ model is shifted to smaller values by 0.3 Å (potentially indicating a slightly too strong bias on β sheets), while all other model distributions are identical to the data distribution. All model distributions of the backbone dihedral angles show the same features as their distribution in the training set with high occurrences in the α-helical, extended, and left-handed helical regions, although the model distributions tend to be more diffuse, spanning a larger area of the Ramachandran map than in the distribution of the training set. These differences reflect the residual KL divergence between the optimized model distribution and the data distribution, arising from the mapping entropy (i.e., that several configurations in the atomistic model translate to the same CG configuration), which is the same for all models, and from the differences in the potential energy functions, which are unable to perfectly describe the native data distribution. For example, the CG protein model employed here allows for slightly more flexibility of the backbone by its side-chain beads filling less space than the full side chains in an atomistic representation, and this manifests in the more diffuse Ramachandran plots of the backbone dihedral angles.

The transferability of the protein models was investigated using a test set of proteins independent of the training set, consisting of all 78 proteins in the ASTRAL database with a SPACI score between 0.7 and 0.8. The data and model distributions of the above-mentioned structural observables were calculated for this test set (see Figures S4 (right), S6, S7 (right), and S8 (right) in the Supporting Information). These model distributions were practically identical to the data distributions of the test set, indicating the transferability of the protein models to proteins not in the training set. This is an improvement over the nontransferable protein model of Winther et al.,[42] who were limited to a small set of short peptides as their training set by the cumbersome calculation of the ensemble averages in the model distributions at every iteration. Here (and in our previous work[20,21]), it is the efficient estimation of the gradient of the logarithm of likelihood by the CD approximation that allows for the employment of a more-realistic training set. We note that other efficient methods also exist to avoid the re-evaluation of ensemble averages (for example, Shell et al. used a reweighting of ensemble averages[12]).

3.3. Accessible Regions of the Ramachandran Plot from MC Simulations of an Ala_{16} Peptide.

To test the available regions of the Ramachandran plot using the two vdW models described in the Methods section, MC simulations of a 16-residue peptide, Ala_{16}, were carried out at room temperature, using the vdW and hydrogen-bond energy contributions, together with the C_α valence backbone stress, without the secondary structure bias. For all models investigated, the accessible regions of the Ramachandran maps in the MC simulations at room temperature cover the allowed regions calculated from the ASTRAL 1.75 database (see Figure S9 in the Supporting Information). On the individual residue level, for all models, helical backbone dihedral angles occur most frequently, with the extended and left helical conformations also being significant. The distributions for the LJ_{learnt} and hard cutoff models are more diffuse and more connected between the positive and negative ϕ values, indicating a smaller energy barrier for the conformational changes of the peptide backbone within these regions of the probability map. During the simulations, there is approximately one hydrogen bond per configuration at any time, indicating that random coil is the main conformation. The hydrogen-bond distribution is plotted in Figure 2. For the hard cutoff model and the LJ model with learnt vdW parameters (LJ_{learnt}), the most commonly observed hydrogen bonds correspond to 3,10- ($i\rightarrow(i-3)$ hydrogen bonds) and α-helices ($i\rightarrow(i-4)$ hydrogen bonds). This is consistent with experimental studies of polypeptides with high alanine content.[69] However, when using the LJ potential with vdW parameters adopted from CHARMM (LJ_{CHARMM}) or AMBER (LJ_{AMBER}), π-helices ($i\rightarrow(i-5)$ hydrogen bonds) are also found to be common, which are not seen experimentally. This problem was also seen in previous molecular dynamics simulations of short peptides[70] using the CHARMM force field. The difference between the hydrogen-bond distribution using the various LJ potential parameters implies that it is possible to change the relative stability of the different helix types by tuning the LJ potential parameters, and this is confirmed by simulations using the hydrogen-bond and the C_α valence angle stress parameters of the LJ_{learnt} model with the LJ parameters of the three LJ models investigated (see Figure S10 in the Supporting Information). For all models, left handed helices ($i\rightarrow i+3,4$) are also present, in agreement with the allowed regions of the Ramachandran map, indicating that turn formation in unbiased loop regions of proteins is conformationally accessible.

3.4. Studying Steric Effects in Turn Conformations on 16-Residue Peptides with a Hairpin Bias.

The protein model employed here is designed to be used with a known (or predicted) secondary structure and $\beta-\beta$ residue contact bias. To further test how the hard cutoff and LJ type vdW models perform in unbiased regions of proteins, in particular in turn regions of β-hairpins, nested sampling simulations of 16-residue peptides were performed employing a hairpin bias, where the turn is located at the center of the peptide (residues 8 and 9). The peptides used in this test were an Ala_{16} peptide, and its mutated forms, where one of the turn residues is replaced by Gly. These will be referred to as A-G-A-A, A-A-G-A and A-A-A-G, corresponding to the glycine being at the i+1, i+2, or i+3 position of the turn, respectively. The secondary structure bias of the energy function keeps the backbone of residues 1−7 and 10−16 extended, as well as restraining the $C_\beta-C_\beta$ distances of the interacting amino acid residue pairs of the two strands. The inner two residues of the turn are unbiased, thus allowing the investigation of whether or not the protein models described in the Methods section reproduce observed correlations between the position of glycine in a β-turn and the observed turn conformation.[71] The turn types found in the NS simulations are listed in Table 3, with their relative probabilities at 298 K, where we used the turn definitions of Venkatachalam[72] (see Figure S11 in the Supporting Information). The relative probability of a turn type at 298 K is calculated by summing the posterior weights of all NS configurations that fall into the definition of the turn type, and then normalizing it by the sum of the posterior weights of all turn types. [The posterior weights of NS configurations are proportional to the available phase space volume at a given temperature; hence, they provide the probability of finding the system in that configuration.] Turn type IV, that is, when no particular turn type can be assigned to the dihedral angles of residues 8 and 9, is omitted from this analysis.

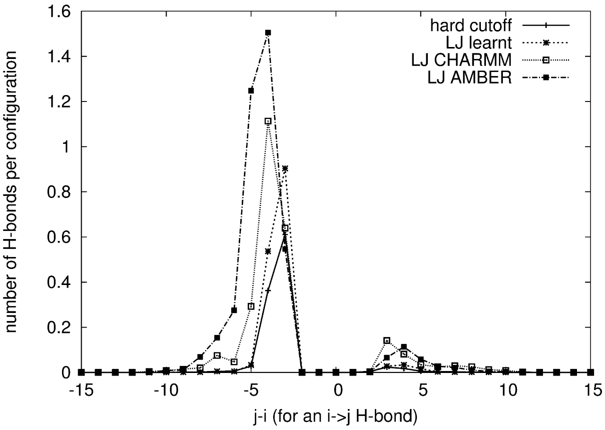


Figure 2. Hydrogen-bond pattern from MC simulations of an Ala_{16} peptide, using the protein models employing the hard cutoff vdW potential (solid line), the LJ_{learnt} model (dashed line), the LJ_{CHARMM} model (dotted line), and the LJ_{AMBER} model (dash-dotted line). Potential parameters are listed in Table 2. On the horizontal axis, −4 represents a hydrogen-bond between amino acid residues $i\rightarrow j = i-4$, typical of α-helices, while −3 is typical of (3,10)-helices, and −5 of π-helices. The small peak between +3 and +5 corresponds to left-handed helices.

All models investigated show the same trend of the turn types adopted in the corresponding simulations, although significant differences between the models used can be observed for simulations of peptides with the i+2 residue of the turn substituted with a glycine (A-A-G-A). In all the

Table 3. Relative Probabilities of the Turn Types Identified from Nested Sampling Simulations of 16-Residue Peptides Applying a β-Hairpin Bias, at 298 K[a]

turn residues	vdW model	turn II′	turn I′	turn I	turn II
AAAA	hard cutoff	0.968	0.000	0.000	0.032
AAAA	LJ_{learnt}	0.983	0.000	0.003	0.014
AAAA	LJ_{CHARMM}	0.965	0.000	0.028	0.000
AAAA	LJ_{AMBER}	0.997	0.000	0.002	0.001
AGAA	hard cutoff	0.980	0.000	0.001	0.020
AGAA	LJ_{learnt}	0.993	0.000	0.001	0.006
AGAA	LJ_{CHARMM}	0.997	0.000	0.003	0.000
AGAA	LJ_{AMBER}	1.000	0.000	0.000	0.000
AAGA	hard cutoff	0.864	0.022	0.001	0.113
AAGA	LJ_{learnt}	0.873	0.023	0.001	0.102
AAGA	LJ_{CHARMM}	0.619	0.091	0.029	0.182
AAGA	LJ_{AMBER}	0.588	0.383	0.001	0.025
AAAG	hard cutoff	0.944	0.000	0.009	0.046
AAAG	LJ_{learnt}	0.980	0.000	0.007	0.012
AAAG	LJ_{CHARMM}	0.931	0.000	0.066	0.000
AAAG	LJ_{AMBER}	0.969	0.000	0.030	0.000

[a] Turn type IV was excluded from the analysis. Substituting the i+1, i+2, or i+3 residue of the turn by glycine (AGAA, AAGA, and AAAG, respectively) increases the relative probability of the type II′, the types I′ and II, and the type I turn, respectively.

simulations of the peptides, the type II′ turn is the dominant turn type. When substituting the i+1 residue of the turn of the polyalanine peptide with a glycine (A-G-A-A), the posterior weight of type II′ turn increases further, and becomes almost the exclusive turn type. This is consistent with the findings of Sibanda et al.[71] that, among the protein structures investigated, type II′ turns mostly occurred with X-G-[ST]-X turn residues (with X being an unspecified amino acid). Substituting the i+3 residue of the turn with a glycine (A-A-A-G) increases the probability of adopting a type I turn (by more than a factor of 2). This is consistent with type I turns typically having glycine residues at the i+3 position of the turn (X-X-X-G).[71] When substituting the $i + 2$ residue of the turn with a glycine (A-A-G-A), the type I′ and type II turns become much more significant compared to simulations of other glycine-substituted peptides. The increase in the probability of type I′ turns is consistent with type I′ turns most often consisting of X-[NDG]-G-X residues.[71] Simulations using the LJ_{AMBER} model appear to demonstrate this best. However, this discrepancy might also be attributable to the LJ_{AMBER} model being best at artificially compensating for the lack of explicit side-chain−main-chain hydrogen bonds in our model. If this were the case, including side-chain-main-chain interactions in our model would further increase the probability of the type I′ turn for an A-[ND]-G-A peptide for the other models investigated, and the apparent advantage of the LJ_{AMBER} model would be lost; however, this is beyond the scope of the present paper.

When comparing the fully learnt (LJ_{learnt}) model with the hard cutoff model, the two models perform very similarly, and consistently with findings in the literature. We find no apparent superiority of the more-elaborate LJ function of the vdW potential in this test. However, note that this does not imply that, generally, vdW interactions would be unimportant in modeling small peptides; for example, they have been found to have a stabilizing effect in quantum mechanical studies of short polyalanine helices.[73] In our CG model, secondary structure bias contributions are optimized to stabilize the secondary structure, and, for this particular model, no superiority of any one of the investigated vdW models is indicated.

3.5. Heat Capacity Curves of an Ala_{16} Peptide with Varying Secondary Structure Bias.

Since purely structural properties of polyalanine peptides are not sufficient to rank the protein models, we also investigated the energetics of the models. However, analyzing the energetics of solely the vdW contributions would be misleading, since all other model parameters might depend on the values of the vdW parameters. Instead, we investigated relative stabilities and heat-capacity curves from polyalanine simulations. One of the major advantages of nested sampling is that, by post-processing the results of the simulation, thermodynamic properties such as heat capacity curves may be calculated for any temperature. Here, we calculate heat capacity curves for a 16-residue polyalanine peptide under the assumption of either an α-helix or β-hairpin secondary structure by using an α-helical or β-hairpin secondary structure bias.

The critical temperatures of the heat-capacity curves (T_c) (i.e., the peak position) and the heat capacities $C_{v,c}$ at these temperatures are listed in Table 4, with the heat capacity curves given in Figure S12 in the Supporting Information. Also shown in Table 4 are some indicative experimental values taken from calorimetric measurements of a variety of peptides 20−30 amino acid residues in length,[74] although the secondary structures of these peptides were not reported. Specific β-hairpin peptides (see, e.g., ref 75) involve a significant amount of stabilizing side-chain interactions which are not modeled by the polyalanine peptides, so they were omitted from this comparison. The heat capacities for all four models correlate better with the experimental values under the assumption of a α-helix rather than a β-hairpin. This is consistent with experimental NMR studies of polyalanine peptides, which find a helical form at room temperature,[69] and strongly suggests that the α-helix form is indeed the more stable.

Table 4. Critical Temperatures (T_c) of Heat-Capacity Curves and the Heat-Capacity Value at T_c ($C_{v,c}$) in Units of R for the Ala_{16} Nested Simulations with α-Helix and β-Hairpin Secondary Structure Bias, Using the Hard Cutoff (Hard) and Lennard-Jones Type vdW Models[a]

	Critical Temperature Data (°C)				
	T_c^{hard}	$T_c^{LJ_{learnt}}$	$T_c^{LJ_{CHARMM}}$	$T_c^{LJ_{AMBER}}$	T_c^{exp}
α-helix	130	70	0	150	0−30
β-hairpin	10	40	20	30	
	Heat-Capacity Data (R)				
	$C_{v,c}^{hard}$	$C_{v,c}^{LJ_{learnt}}$	$C_{v,c}^{LJ_{CHARMM}}$	$C_{v,c}^{LJ_{AMBER}}$	$C_{v,c}^{exp}$
α-helix	170	130	90	80	100−200
β-hairpin	67	63	43	47	

[a] Approximate experimental values (exp) are taken from ref 74.

Of the four models, the LJ_{CHARMM} model initially appears to give the best prediction for the critical temperature. However, this is the only simulation that predicts the β-hairpin to be more stable than the α-helix (i.e., to have a higher T_c values). In contrast, the very high critical temperatures predicted for the hard cutoff potential and LJ_{AMBER} model show that these models cause the α-helix secondary structure to be overly stable, which is consistent with the critical temperature (~400 K, or 127 °C) found by Peng et al.,[76] using the AMBER force field for a 15-residue polyalanine peptide. The critical

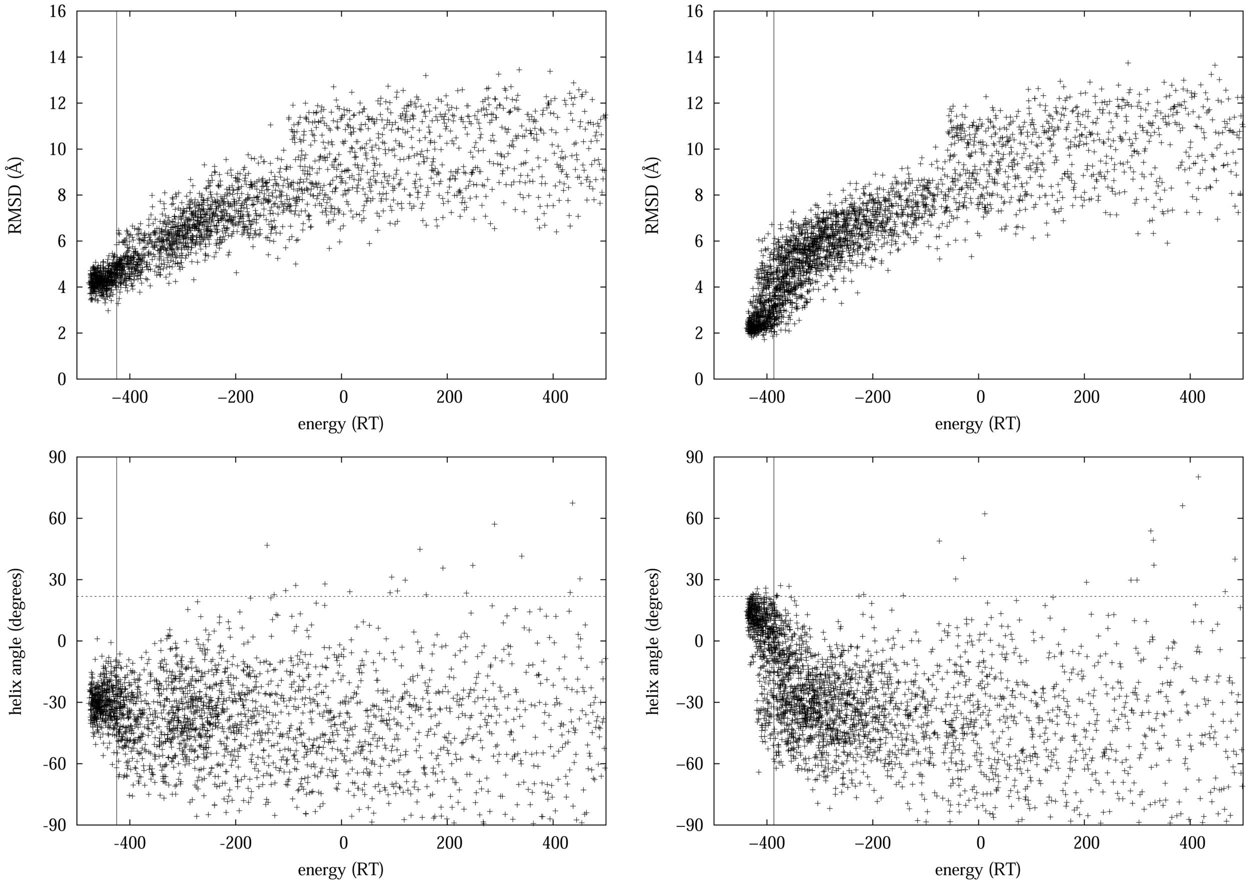


Figure 3. The backbone RMSD from the native state (top), and the angle of the helix with respect to the axis of the β-strands (bottom), as a function of the potential energy for the conformations in the main basin of the energy landscape, explored by nested sampling simulations using the protein model with (left) hard cutoff vdW potential and (right) Lennard-Jones type vdW potential with inferred vdW parameters. The estimated energy at room temperature is marked by solid vertical lines. Conformations obtained by using the LJ potential show a wide range of allowed helix orientation angle at room temperature, including the native angle in the crystal structure, 21.8° (dashed horizontal line), while simulations using the hard cutoff potential fail to find the native helix orientation.

temperatures calculated using the $\mathrm{LJ}_{\mathrm{learnt}}$ model correlate best with the experimentally observed range while still predicting the correct relative secondary structure stability. Comparing the hard cutoff model and the $\mathrm{LJ}_{\mathrm{learnt}}$ model (for which all parameters including the LJ parameters were inferred), the heat capacity results and the relative stabilities suggest that the LJ potential may be more suitable than the hard cutoff vdW potential for calculating the thermodynamic properties of peptides. We also note that there is sufficient flexibility in the LJ parameters to change the relative stabilities of helical and strand conformations. The LJ parameter values of the $\mathrm{LJ}_{\mathrm{CHARMM}}$ model overstabilizes the β-hairpin form, while the $\mathrm{LJ}_{\mathrm{AMBER}}$ model overstabilizes the α-helical form, when used with the CRANKITE force field.

3.6. Folding Simulations of Protein G. In the previous sections, simulations of small peptides with fixed secondary structure were described, where the form of the vdW interactions was found to have no significance using our CG force field including a secondary structure bias. The effects of the long-range vdW interaction contributions of our force field on side-chain packing between interacting secondary structure units (α-helices and β-sheets) can be investigated by protein folding simulations, because this tertiary level of structure formation is not modeled by other interaction parameters in our force field. We present folding simulations of protein G, including secondary structure bias and hydrophobic interaction contributions in the models used. Protein G is a 56-residue protein consisting of an antiparallel four-stranded β-sheet and an α-helix, with a β-Grasp (ubiquitin-like) fold (see Figure S13 (right) in the Supporting Information). Conformations found in simulations using the different vdW models were assessed visually (which side of the β-sheet the helix was on, whether the hydrophobic residues are in the interior of the protein or exposed), as well as quantitatively, by calculating the C_α root-mean-square distance (RMSD) from the crystal structure present in the PDB database, and the angle of the helix orientation with respect to the axis of the β-sheet. The helix orientation angle is calculated as the directional angle between the axis of the N-terminal β-strand (the vector pointing from the C_α atom of residue 7 to the C_α atom of residue 3) and the axis of the α-helix (the vector pointing from the center of mass of the C_α atoms of residues 24−27 to the center of mass of the C_α atoms of residues 31−34), around the surface normal of the β-sheet (the cross product of the vector pointing from the C_α atom of residue 7 to the C_α atom of residue 3, and the vector

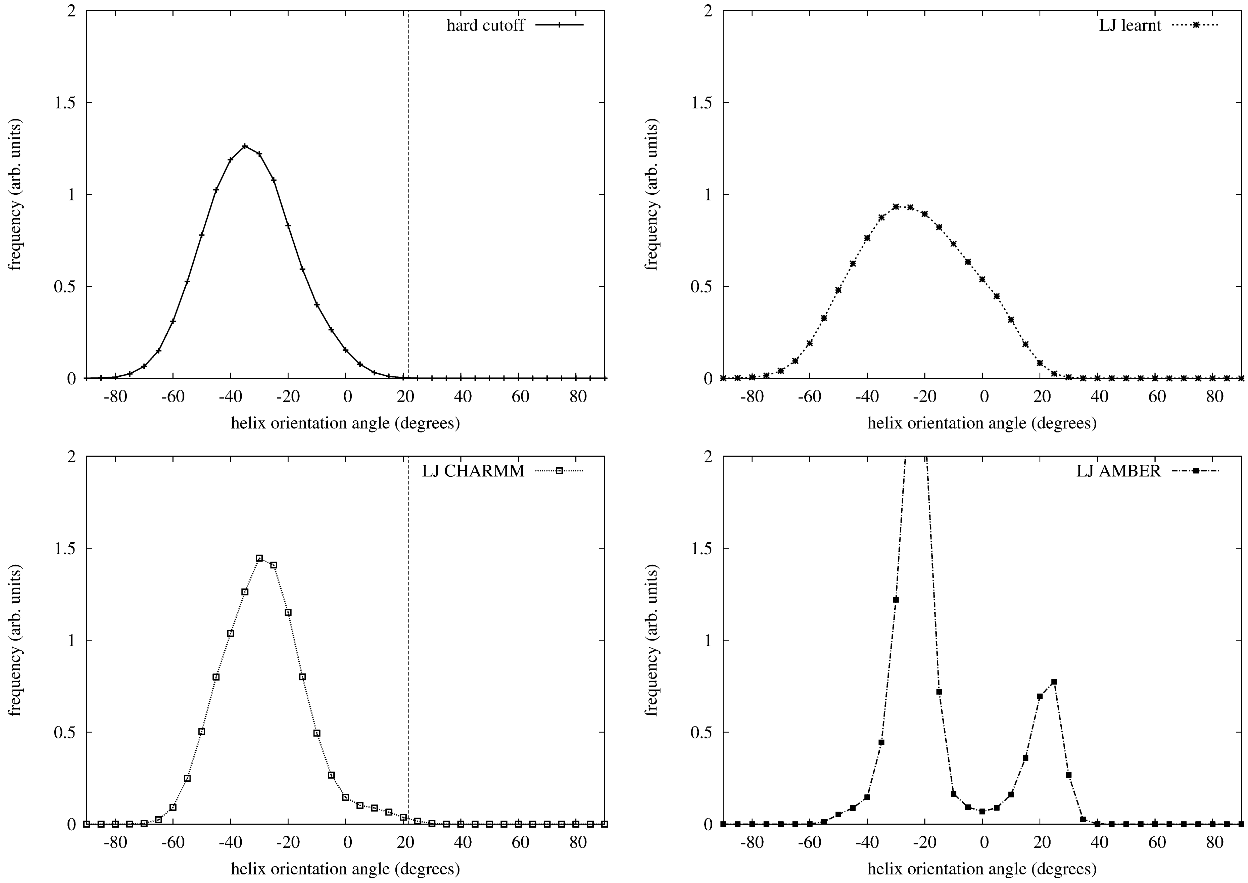


Figure 4. Distribution of the helix angle at room temperature from a MC simulation for the different models: (top left) hard cutoff model, (top right) LJ_{learnt}, (bottom left) LJ_{CHARMM}, and (bottom right) LJ_{AMBER}. Simulation length: 10^{10} MC steps, starting from the crystal structure. Vertical dashed lines show the helix orientation angle in the crystal structure.

pointing from the C_α atom of residue 7 to the C_α atom of residue 54).

For all vdW models investigated here, the main conformation at room temperature is topologically correct. The helix was on the correct side of the β-sheet at room temperature in all simulations, as opposed to earlier simulations using the CRANKITE protein model without including the γ atoms (and without hydrophobic interactions), which allowed the helix to be equally on either side of the sheet.[21] Since there is no information coded in the secondary structure bias about which side of the sheet the helix may pack against, this indicates that having the γ atoms and the hydrophobic interactions in the model makes a clear distinction between the two basins. Previous simulations including γ atoms but no hydrophobic interactions (data not shown) showed a preference for the helix to be on the correct side of the sheet, probably due to the steric clashes of large residues in the loops that prohibit the folding of the helix onto the wrong side of the sheet at room temperature. The inclusion of hydrophobic interactions enables a qualitative shaping of the energy landscape, representing a driving force for the correct collapse of the protein in the folding simulations, in agreement with previous studies arguing for the importance of the hydrophobic interactions in protein folding.[77]

When comparing the RMSD of the conformations in the main basin from the native conformation in the PDB database, the LJ_{learnt} model outperforms the hard cutoff potential. Conformations in the main basin of the energy landscape, explored by NS simulations using the LJ_{learnt} model, have an RMSD from the native conformation as small as 2 Å, while the model employing the hard cutoff potential cannot find conformers that have an RMSD distance of less than 3 Å (see Figure 3, top). The reason for this is that the packing of the helix with respect to the β sheet can be better described by the LJ model. Indeed, the orientation of the α-helix, with respect to the β-sheet, is closer to the native orientation when using the LJ_{learnt} model (see Figure 3 (bottom), as well as Figure S13 in the Supporting Information). The native helix orientation angle, with respect to the sheet, only appears using the LJ potential, and a wide range of orientation angles are accessible at room temperature, showing that a twisting motion of the helix is allowed. This is consistent with rigidity analysis of Protein G,[22] where the lowest-frequency nontrivial mode of the normal-mode analysis of Protein G was found to correspond to a rotation of the helix about an axis perpendicular to the β-sheet, allowing a deviation of more than 30° in the helix orientation angle from the crystal structure while maintaining the network of hydrophobic bonds present in the crystal structure.

The reasons why the LJ potential form could be better than the hard cutoff at modeling the packing of Protein G could be 2-fold. First, as discussed in section 3.2, the LJ potential is softer than the hard cutoff potential, allowing for more flexibility of the loop regions at the two ends of the helix; and second, the weak long-range attractive interactions might favor the packing of the helix in the native orientation, which would appear as a zero-energy contribution using a hard cutoff. However, we have found that the hard cutoff and the LJ potentials behaved similarly in modeling small loop regions of peptides with simple tertiary structure, suggesting that it is more likely that the long-range attractive interactions make the Lennard-Jones potential a more-realistic model for proteins. Our results confirm previous observations about the importance of the long-range attractive interactions of the vdW interactions in the modeling of the packing of protein interior[78] and small clusters.[79] We find that, in the CRANKITE model, while the hydrophobic interactions are responsible for stabilizing the correct tertiary assembly of the secondary structure elements enabling the qualitatively correct collapse of the protein during the folding process, the vdW interactions are important for the fine-tuning of the energy landscape within its main basin. This agrees with previous experimental and simulation results (see citations given in ref 80), which found that both the hydrophobic interactions and the packing are important in protein folding.

When comparing simulations using the LJ potential with learnt or adopted vdW parameters, we find that, although low RMSD structures with the native orientation are observed in all LJ simulations (see Figure S14 in the Supporting Information), the distributions of the helix orientation angle exhibit significant differences: while the helix distribution angle follows a broad unimodal distribution for the $\mathrm{LJ}_{\mathrm{learnt}}$ model, it follows a bimodal distribution using the $\mathrm{LJ}_{\mathrm{CHARMM}}$ and $\mathrm{LJ}_{\mathrm{AMBER}}$ models, implying a two-state model with a high energy barrier. This is shown by the distribution (Figure 4), the trace plots (Figure S15 in the Supporting Information), and the autocorrelation functions (Figure S16 in the Supporting Information) of the helix orientation angle, calculated in room temperature MC simulations of 10^{10} steps, starting from the crystal structure. The energy barrier of twisting the helix is so high using the $\mathrm{LJ}_{\mathrm{AMBER}}$ model that the helix orientation angle only switched once between the two main basins. The trace plots and the long autocorrelation time of the helix orientation angle of the $\mathrm{LJ}_{\mathrm{CHARMM}}$ model suggest the presence of an energy barrier for this model. The rigidity analysis of Protein G[22] suggests a broad unimodal distribution without the implication of an energy barrier, supporting the distribution generated by the $\mathrm{LJ}_{\mathrm{learnt}}$ model. We note that the helix angle distribution is far from perfect, being shifted toward negative values, which indicates that there are other effects not considered in the model that play a role in the helix packing, for example, electrostatic interactions.

4. DISCUSSION

When inferring a generalizable protein force field using a training set of proteins with varying sequences (see section 2.1), our ML approach with the CD approximation relies on the following assumptions. First, the protein conformations of the various sequences S_0^i come from their respective Boltzmann distributions corresponding to the same inverse temperature, and second, the training set of protein conformations represents independent and representative samples from a set of proteins that is intended to be modeled by the protein force field.

The training set of protein conformations may be experimentally observed,[20,21,42] or computer-generated.[12,28] When conformations are generated from computer simulations at a given temperature, although the assumption of Boltzmann distribution of each sequence holds a priori, the fitted CG model will have the limitations of the all-atom model at best. The same holds for fitting to NMR structures optimized by all-atom force fields. Hence, we used only crystal structures in the training set of our protein model. The assumption that the individual conformations in the training set, all of which are crystal structures, are representative of the native structure in thermodynamic equilibrium in solution, is based on previous studies.[81,82] When the atomic coordinates of proteins are mapped onto the CG model, high energy states, non-representative of the Boltzmann distribution were eliminated by removing the clashing gamma atoms. This causes the converged parameter values (hydrogen bond strength, bias potential strength) to be consistently up to 5% lower than when the ensemble including high energy conformations is used (Figure 1a). A possible explanation of this is that stronger attractive interactions (hydrogen bonds and side-chain−side-chain interaction) are necessary to compensate the high energy atomic clashes, in order to be able to preserve the structure of the proteins in an MC simulation. This demonstrates the importance of the data set of known proteins being drawn from an ensemble representing thermodynamic equilibrium at room temperature. One might argue that it could be better to keep all atoms, and relax the structure by minimization or perturbation of the structures. However, at this stage, we do not know the parameters of the energy function, and the energy function used would bias the equilibrium state, and the inferred potential parameter values. We also note that, in PDB structures, there are missing atoms, and none of the potential parameters of our CG model are dependent on whether all atoms in a residue are present. In the parameter inference, the dataset with the clashing atoms removed was used.

According to the Boltzmann hypothesis, the statistics of structural features such as hydrogen-bond distances in the native state of proteins comply with the Boltzmann distribution.[83−85] It has been argued that the Boltzmann hypothesis represents an evolutionary equilibrium where these structural features are maintained around a narrow set of values,[83] for example it has been proposed that protein sequences have evolved maintaining an optimal mean hydrophobicity profile.[84] According to the maximum entropy principle, these may be considered as evolutionary constraints on the evolution of protein sequences (see the Discussion section in the work of Podtelezhnikov et al.[27]). This argument suggests the existence of a generalizable protein force field that captures these evolutionary constraints, which we infer using a training set of protein conformations that is representative of the proteins to be modeled (that is, proteins with a globular structure). In another study, to recover a very simple underlying CG force field, a training set of 5 proteins have been found to be sufficient,[28] where the training set is called an extended canonical ensemble, referring to the collection of equilibrium systems that are governed by the same underlying general force field.

To test that our training set is representative of this distribution, we considered parameter estimation using different subsets of the ASTRAL library, marked by a minimum SPACI score, representing the quality of the crystallographic structures. The higher the SPACI score the better the crystallographic structures are, although the variability of folds may be lower, due to the smaller number of structures. The parameter estimation using the different subsets reveals a trend for the hydrogen bond strength (a 10% increase for SPACI score 0.8 as opposed to 0.4), corresponding to more perfectly formed hydrogen bonds in the dataset, but no dependence of the bias potential parameters on the quality of protein structures. The weak dependence of the protein model parameters on the quality of the crystal structures indicate that the ASTRAL data set is sufficiently diverse to estimate parameters of a generalizable protein model, and as such, in the parameter inference, we used the subset of the database with a minimum SPACI score of 0.8, comprising 73 proteins of varying length from 43 to 690. In comparison, Winther and Krogh[42] used a dataset of 24, 11−14-residue-long peptides as the training set of their ML inference. Although the training set was successfully folded with their optimized potential, the inferred protein model was not found to be transferable to peptides not included in the training set. One of the reasons for this was that the training set was not representative of the native distribution of protein sequences.

The CD approximation allows a significant acceleration of the ML inference. Assuming 10^6 MC steps for the convergence of the ensemble average, which might be a reasonable estimate for the peptide size used by Winter and Krogh,[42] the acceleration of the ML inference coming from the use of the CD approximation is over 200-fold for the same dataset of peptide conformations. Moreover, larger proteins included in the dataset will have longer equilibration and decorrelation times (for example, in the MC simulations of Protein G using the $\mathrm{LJ}_{\mathrm{AMBER}}$ model, even 10^{10} MC steps are not sufficient to calculate the equilibrated distributions), further increasing the acceleration of the current algorithm over a naïve ML algorithm.

5. CONCLUSION

In this work, the potential parameters of a generalizable coarse-grained (CG) force field for modeling proteins were inferred, or learnt, from a data set of known protein structures, using a maximum likelihood (ML) approach. We show how our method of inferring a generalizable protein model relates to inferring protein models specific to an amino acid sequence. This ML inference of a specific force field relies on the assumption that the training set contains independent observations of conformations of not only one, but a set of proteins, which are independent and representative of the proteins to be modeled by the force field. While the training set used here is a subset of crystal structures from the Protein Database (PDB) database (the only available experimental data on protein structures), it could also be generated by computer simulations.[12,28]

To avoid the necessity of equilibrating each protein of the training set in the model distribution at each iteration of the ML optimization, we employ contrastive divergence for a computationally efficient approximation of the gradient of the energy with respect to the potential parameters, reducing the computational requirements by several orders of magnitude. The contrastive divergence approximation relies on the assumption that the conformations of any protein in the training set represent samples from a thermal equilibrium. We show that if this assumption does not hold (due to including several high energy conformations), a systematic error in the parameter estimation is introduced. The algorithm is very simple, increasing the number of the parameters of the ML inference by only two; the number and the maximum amplitude of Monte Carlo (MC) steps to generate the perturbed data distribution. While the number of MC steps only affects the noise on the gradient estimate, we find that, because of the ruggedness of the energy landscape, selection of the maximum allowed MC step size affects the local exploration of the energy landscape. Preliminary tests show that the ML optimization can be further accelerated by employing an adaptive learning rate with an associated momentum, as suggested by Hinton.[51]

We infer parameters for protein models employing two different van der Waals (vdW) interaction potentials: a hard cutoff potential and a Lennard-Jones (LJ) potential using inferred parameters ($\mathrm{LJ}_{\mathrm{learnt}}$) and parameters adopted from the CHARMM and AMBER force fields ($\mathrm{LJ}_{\mathrm{CHARMM}}$ and $\mathrm{LJ}_{\mathrm{AMBER}}$, respectively). We find that the $\mathrm{LJ}_{\mathrm{learnt}}$ model better models heat capacities of small peptides, as well as the helix orientation distribution of Protein G at room temperature, when used within the CRANKITE force field, which is an improvement over the original version of the force field employing the hard cutoff potential form. In the improved force field, the hydrophobic interactions determine the main basin of the energy landscape into which the protein collapses during the folding simulations, while the vdW interactions serve to fine-tune the potential energy landscape within the main basin. The simulation results suggest that the CRANKITE force field can be further improved by incorporating electrostatic interactions or side-chain−main-chain hydrogen-bond interactions. Our simulations demonstrate that model parameters generally are not transferable between different models. When comparing the all-atom CHARMM or AMBER force fields using our CG force field, both the atomistic resolution and the energy function differ significantly. Adopting vdW parameters without further optimization was found to cause a significant change in the secondary structure bias potential parameters (not present in the CHARMM or AMBER force fields), and the relative stability of the secondary structure elements was also found to be altered. However, the maximum likelihood inference using the contrastive divergence approximation employed here provides an efficient general inference scheme to achieve a model distribution closest to the data distribution in the training set, as long as the assumptions of the model discussed above hold.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information contains details of the function form of the hydrogen-bond interaction, the convergence plot of the hydrogen-bond strength parameter and the distribution of the $\nabla_\theta \log L$ estimate, the LJ potential form (as implemented in this work), the acceptance rate (as a function of the MC step size), distribution plots of backbone dihedral angles, α carbon valence angles and β carbon distances in β-sheet interactions for both the training set of proteins and the independent test set of proteins, the distribution plots of backbone dihedral angles and H-bond interactions for the Ala_{16} peptides, the turn type definitions, heat capacity curves of Ala_{16} peptides, the conformational ensembles at room temperature from NS simulations of Protein G for the hard cutoff and $\mathrm{LJ}_{\mathrm{learnt}}$ potential, trace plots and autocorrelation functions of MC simulations of Protein G, and a list of the training set of proteins. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: D.L.Wild@warwick.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge support from the Leverhulme Trust (Grant No. F/00 215/BL). We thank Dr. Stephen A. Wells for helpful discussions on the rigidity analysis of Protein G.

■ REFERENCES

(1) http://www.predictioncenter.org/.
(2) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res.* **2005**, *33*, D501−D504.
(3) *Nat. New Biol.* **1971**, *233*, 223.
(4) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Protein Sci.* **2011**, *334*, 517−520.
(5) Takada, S. *Curr. Opin. Struct. Biol.* **2012**, *22*, 130−137.
(6) Tozzini, V. *Q. Rev. Biophys.* **2010**, *43*, 333−371.
(7) Tirion, M. M. *Phys. Rev. Lett.* **1983**, *80*, 3696−3700.
(8) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. *Chem. Rev.* **2010**, *110*, 1463−1497.
(9) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183−210.
(10) Nguyen, H. D.; Hall, C. K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16180−16185.
(11) Bereau, T.; Desetno, M. *J. Chem. Phys.* **2009**, *130*, 235106.
(12) Carmichael, S. P.; Shell, M. S. *J. Phys. Chem. B* **2012**, *116*, 8383−8393.
(13) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849−873.
(14) Sorenson, J. M.; Head-Gordon, T. *Prot. Struct. Funct. Gen.* **2002**, *46*, 368−379.
(15) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147−155.
(16) Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. *J. Phys. Chem. B* **2009**, *113*, 267−274.
(17) Irbäck, A.; Sjunnesson, F.; Wallin, S. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13614−13618.
(18) Chen, N.-Y.; Su, Z.-Y.; Mou, C.-Y. *Phys. Rev. Lett.* **2006**, *96*, 078103(1−4).
(19) Podtelezhnikov, A. A.; Wild, D. L. *Source Code Biol. Med.* **2008**, *3*, 12.
(20) Podtelezhnikov, A. A.; Ghahramani, Z.; Wild, D. L. *Prot. Struct. Funct. Bioinf.* **2007**, *66*, 588−99.
(21) Podtelezhnikov, A. A.; Wild, D. L. *Biophys. J.* **2009**, *96*, 4399−4408.
(22) Burkoff, N. S.; Várnai, C.; Wells, S. A.; Wild, D. L. *Biophys. J.* **2012**, *102*, 878−886.
(23) Burkoff, N. S.; Várnai, C.; Wild, D. L. *Bioinformatics* **2013**, *29*, 580−587.
(24) Moore, W. J. *Physical Chemistry*, 4th Edition; Prentice−Hall, Inc: Englewood Cliffs, NJ, 1972; pp 617−644.
(25) Cossio, P.; Trovato, A.; Petrucci, F.; Seno, F.; Maritan, A.; Laio, A. *PLOS Comput. Biol.* **2010**, *6*, e1000957.
(26) Anfinsen, C. *Science* **1973**, *181*, 223−230.
(27) Podtelezhnikov, A. A.; Wild, D. L. In *Bayesian Methods in Structural Bioinformatics*; Hamelryck, T., Mardia, K., Ferkinghoff-Borg, J., Eds.; Springer−Verlag: Berlin, Heidelberg, 2012; Chapter 5, pp 135−143.
(28) Mullinax, J. W.; Noid, W. G. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19867−19872.
(29) Hamelryck, T.; Borg, M.; Paluszewski, M.; Paulsen, J.; Frellsen, J.; Andreetta, C.; Boomsma, W.; Bottaro, S.; Ferkinghoff-Borg, J. *PLOS ONE* **2010**, *5*, e13714.
(30) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457−469.
(31) Borg, M.; Ferkinghoff-Borg, T. H. J. In *Bayesian Methods in Structural Bioinformatics*; Hamelryck, T., Mardia, K., Ferkinghoff-Borg, J., Eds.; Springer−Verlag: Berlin, Heidelberg, 2012; Chapter 3, pp 97−124.
(32) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins: Struct. Funct. Bioinf.* **2007**, *69*, 394−408.
(33) Fujitsuka, Y.; Luthey-Schulten, S. T. Z. A.; Wolynes, P. G. *Proteins: Struct. Func. Bioinf.* **2004**, *54*, 88−103.
(34) Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16934−16949.
(35) Vendruscolo, M.; Domany, E. *J. Chem. Phys.* **1998**, *109*, 11101−11108.
(36) Hu, C.; Li, X.; Liang, J. *Bioinformatics* **2004**, *20*, 3080−3098.
(37) Maiorov, V. N.; Crippen, G. M. *J. Mol. Biol.* **1992**, *227*, 876−888.
(38) Mourik, J. V.; Clementi, C.; Maritan, A.; Seno, F.; Banavar, J. R. *J. Chem. Phys.* **1999**, *110*, 10123.
(39) Hao, M. H.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984−4989.
(40) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918−4922.
(41) Ming, D.; Wall, M. E. *Phys. Rev. Lett.* **2005**, *95*, 198201-1−198201-4.
(42) Winther, O.; Krogh, A. *Phys. Rev. E* **2004**, *70*, 030903.
(43) Kleinman, C. L.; Rodrigue, N.; Bonnard, C.; Philippe, H.; Lartillot, N. *BMC Bioinf.* **2006**, *7*, 326.
(44) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.
(45) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.
(46) Chaimovich, A.; Shell, M. S. *J. Chem. Phys.* **2011**, *134*, 094111-1−09411-12.
(47) Rudzinski, J. F.; Noid, W. G. *J. Chem. Phys.* **2011**, *135*, 214101-1−214101-15.
(48) Hinton, G. E. *Neural Computation* **2002**, *14*, 1771−1800.
(49) Hinton, G. E.; Sejnowski, T. J. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundation*; Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, MA, 1986; Chapter 7, pp 282−317.
(50) Smolensky, P. In *Parallel Distributed Computing: Explorations in the Microstructure of Cognition.*; Rumelhart, D. E., MacClelland, J. L., Eds.; MIT Press: Cambridge, MA, 1986; Vol. *1*; pp 194−281.
(51) Hinton, G. *A Practical Guide to Training Restricted Boltzmann Machines*, Technical Report UTML TR 2010-003, University of Toronto, Toronto, Canada, 2010
(52) Bilionis, I.; Zabaras, N. *J. Chem. Phys.* **2013**, *138*, 044313-1−044313-12.
(53) Lennard-Jones, J. *Proc. R. Soc. A* **1924**, *106A*, 441.
(54) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A..Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross,W. S.; Kollman, P. A. *AMBER 9*; University of California, San Francisco, CA, 2006.
(55) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(56) Srinivasan, R.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14258−14263.
(57) Podtelezhnikov, A. A.; Wild, D. L. *Proteins: Struct. Funct. Bioinf.* **2005**, *61*, 94−104.
(58) Engh, R. A.; Huber, R. In *International Tables for Crystallography*, 1st ed.; Rossman, M. G., Arnold, E., Eds.; Kluwer Academic Publishers for the International Union of Crystallography: Dordrecht, Boston, London, 2001; Vol. *F*; pp 382−392.

(59) Ho, B. K.; Coutsias, E. A.; Seok, C.; Dill, K. A. *Protein Sci.* **2005**, *14*, 1011−1018.

(60) Shimada, J.; Kussell, E. L.; Shakhnovich, E. I. *J. Mol. Biol.* **2001**, *308*, 79−95.

(61) Maisuradze, G. G.; Senet, P.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2010**, *114*, 4471−4485.

(62) Lomize, A. L.; Pogozheva, M. Y. R. I. D. *Protein Sci.* **2002**, *11*, 1984−2000.

(63) Buckingham, R. A. *Proc. R. Soc. A* **1938**, *168*, 264−283.

(64) von Lilienfeld, O. A.; Tkachenko, A. *J. Chem. Phys.* **2010**, *132*, 234109.

(65) Fedorenko, R. P. *USSR Comput. Math. Math. Phys.* **1964**, *4*, 227−235.

(66) Chandonia, J. M.; Hon, G.; Walker, N. S.; Conte, L. L.; Koehl, P.; Brenner, M. L. S. E. *Nucleic Acids Res.* **2004**, *32*, D189−D192.

(67) Brenner, S. E.; Koehl, P.; Levitt, M. *Nucleic Acids Res.* **2002**, *28*, 254−256.

(68) Skilling, J. *J. Bayesian Anal.* **2006**, *1*, 833−860.

(69) Chakrabartty, A.; Schellman, J. A.; Baldwin, R. L. *Nature* **1991**, *351*, 586−588.

(70) Armen, R.; Alonso, D. O. V.; Daggett, V. *Protein Sci.* **2003**, *12*, 1145−1157.

(71) Sibanda, B. C.; Bundell, T. L.; Thornton, J. M. *J. Mol. Biol.* **1989**, *206*, 759−777.

(72) Venkatachalam, C. M. *Biopolymers* **1968**, *6*, 1425−1436.

(73) Tkachenko, A.; Rossi, M.; Blum, V.; Ireta, J.; Scheffler, M. *Phys. Rev. Lett.* **2011**, *106*, 118102.

(74) Richardson, J. M.; Makhatadze, G. I. *J. Mol. Biol.* **2004**, *335*, 1029−1037.

(75) Skwierawska, A.; Oldziej, S.; Liwo, A.; Scheraga, H. A. *Biopolymers* **2009**, *91*, 37−51.

(76) Peng, Y.; Hansmann, U. H. E.; Alves, N. A. *J. Chem. Phys.* **2003**, *118*, 2374−2380.

(77) Dill, K. A. *Biochemistry* **1990**, *29*, 7133−7155.

(78) Lammert, H.; Wolynes, P. G.; Onuchic, J. N. *Proteins: Struct. Funct. Bioinf.* **2012**, *80*, 362−373.

(79) Braier, P. A.; Berry, R. S.; Wales, D. J. *J. Chem. Phys.* **1990**, *93*, 8745.

(80) Baldwin, R. L. *J. Mol. Biol.* **2007**, *371*, 283−301.

(81) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. *Proteins: Struct. Funct. Gen.* **1995**, *23*, 142−150.

(82) Best, R. B.; Lindorff-Larsen, K.; DePristo, M. A.; Vendruscolo, M. *J. Chem. Phys.* **2006**, *103*, 10901−10906.

(83) Shortle, D. *Protein Sci.* **2003**, *12*, 1298−1302.

(84) Bastolla, U.; Porto, M.; Roman, H. E.; Vendruscolo, M. *Gene* **2005**, *347*, 219−230.

(85) Jaynes, E. T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, U.K., 2003; pp 1298−1302.

Supporting information for: Efficient parameter estimation of generalisable coarse grained protein force fields using contrastive divergence: a maximum likelihood approach

C. Várnai, N. S. Burkoff, and D. L. Wild*

Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, United Kingdom

E-mail: d.l.wild@warwick.ac.uk

*To whom correspondence should be addressed

1 The hydrogen bonding energy term ($n^{\mathrm{HB}}_{l\to m}$) of the CRANKITE force field

The hydrogen bonding energy term between the amide H atom of residue l (H_l) and the carbonyl O atom of residue m, (O_m), is determined using a distance cutoff, δ, and two angle cutoffs, Θ_{COH} and Ψ_{OHN}.

$$n^{\mathrm{HB}}_{l\to m} = f^{\mathrm{cut}}_{\delta}(d(\mathrm{O}_m,\mathrm{H}_l),\delta) f^{\mathrm{cut}}_{\Theta}(\Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_m),\Theta_{\mathrm{COH}}) f^{\mathrm{cut}}_{\Psi}(\Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l),\Psi_{\mathrm{OHN}}) \tag{1}$$

where $d(O_m,H_l)$ is the O_m–H_l distance, $\Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_m) \geq \Theta_{\mathrm{COH}}$ is the C_m–O_m–H_l angle, and $\Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l) \geq \Psi_{\mathrm{OHN}}$ is the the O_m–H_l–N_l angle. The cutoff functions are

$$f^{\mathrm{cut}}_{\delta}(d(\mathrm{O}_m,\mathrm{H}_l),\delta) = \begin{cases} 1 & d(O_m,H_l) \leq \delta \\ \delta^4 d(\mathrm{O}_m,\mathrm{H}_l)^{-4} & d(\mathrm{O}_m,\mathrm{H}_l) > \delta \end{cases}, \tag{2}$$

$$f^{\mathrm{cut}}_{\Theta}(\Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_m),\Theta_{\mathrm{COH}}) = \begin{cases} 1 & \Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_m) \geq \Theta_{\mathrm{COH}} \\ \cos^4(\Theta_{\mathrm{COH}})\cos^{-4}(\Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_l)) & \Theta(\mathrm{C}_m,\mathrm{O}_m,\mathrm{H}_l) < \Theta_{\mathrm{COH}} \end{cases} \quad \text{and} \tag{3}$$

$$f^{\mathrm{cut}}_{\Psi}(\Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l),\Psi_{\mathrm{OHN}}) = \begin{cases} 1 & \Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l) \geq \Psi_{\mathrm{OHN}} \\ \cos^4(\Psi_{\mathrm{OHN}})\cos^{-4}(\Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l)) & \Psi(\mathrm{O}_m,\mathrm{H}_l,\mathrm{N}_l) < \Psi_{\mathrm{OHN}} \end{cases}. \tag{4}$$

2 Figures referenced in the Main Text

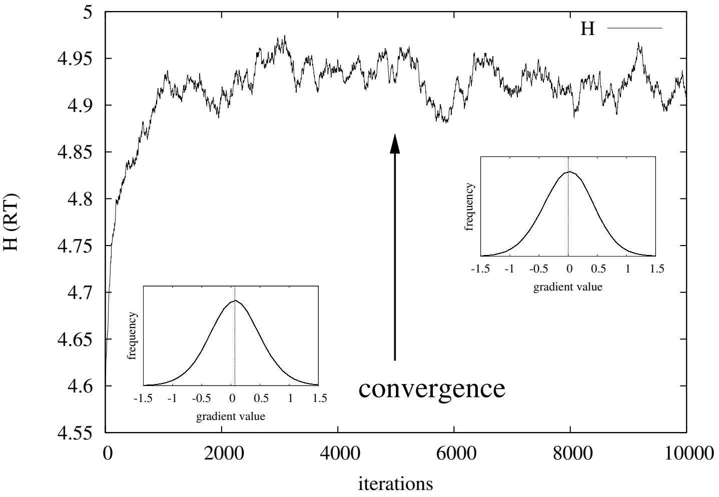


Figure S1: The convergence of the H-bond strength parameter (H) in the ML inference using the CD approximation, when using the hard cutoff model. The insets show the distribution of the approximate gradient for the initial value, $H = 4.25RT$ (left), and for the converged value, $H = 4.92RT$ (right). At the initial value, the distribution is centred around 0.07, while at the converged value, around 0. The standard deviation of the gradient distributions is 0.45. The converged value of the parameter was calculated from iterations 5,000 to 10,000.

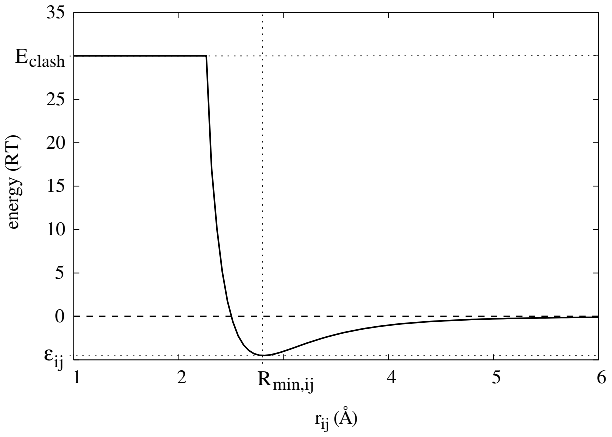


Figure S2: The Lennard-Jones type interaction energy function, as implemented in this work (solid line). The 0 energy is marked with dashed line, the excess energy of clashing atoms (E_{clash}) and the minimum energy (ε_{ij}) are marked with horizontal dotted lines, and the minimum energy separation $R_{\text{min},ij}$ is marked with a vertical dotted line.

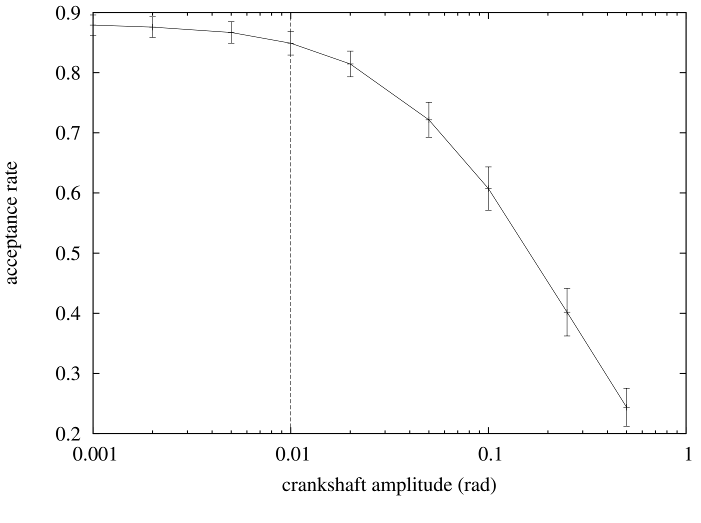


Figure S3: The dependence of the Monte Carlo acceptance rate on the MC step size (solid line). The angle of the maximum allowed crankshaft rotation is on the horizontal axis. The vertical dashed line marks 0.01 radians as the maximum allowed angle of the crankshaft rotations.

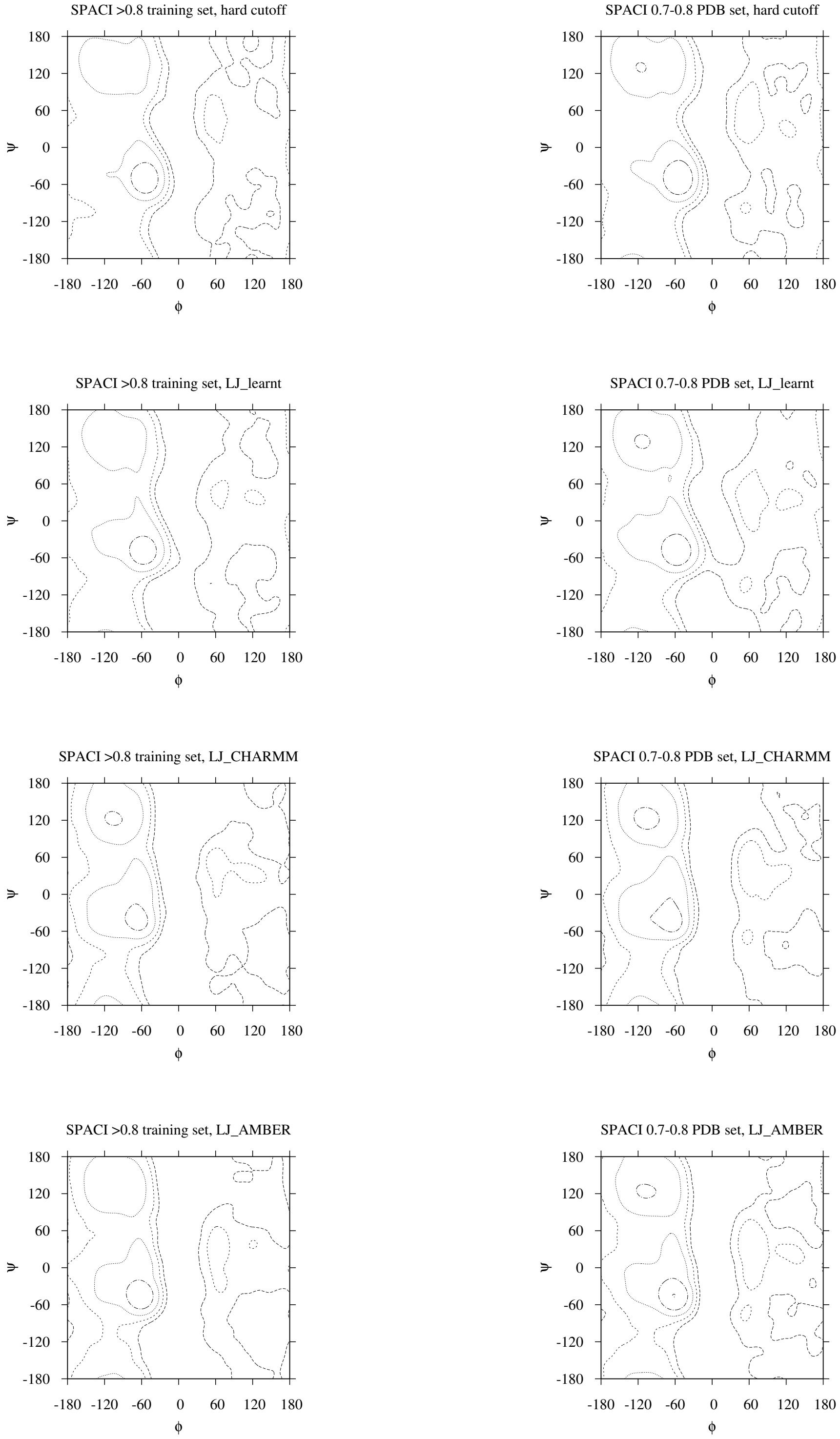


Figure S4: The distribution of backbone dihedral angles in the model distributions using the models listed in Table 1 of the main text (hard cutoff: first row; LJ_{learnt}: second row; LJ_{CHARMM}: third row; LJ_{AMBER}: bottom row). The PDB structures modelled were Left: the training set, Right: an independent PDB set, the ASTRAL PDB structures with SPACI scores from 0.7 to 0.8.

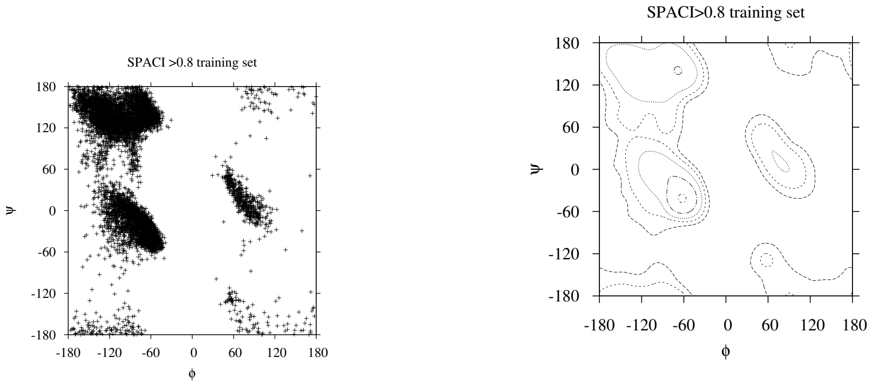


Figure S5: The Ramachandran map of the backbone dihedral angles in the training set (the data distribution). Left: backbone dihedral angles without smoothing, Right: backbone dihedral angles smoothed by Gaussians with standard deviation of 10 degrees.

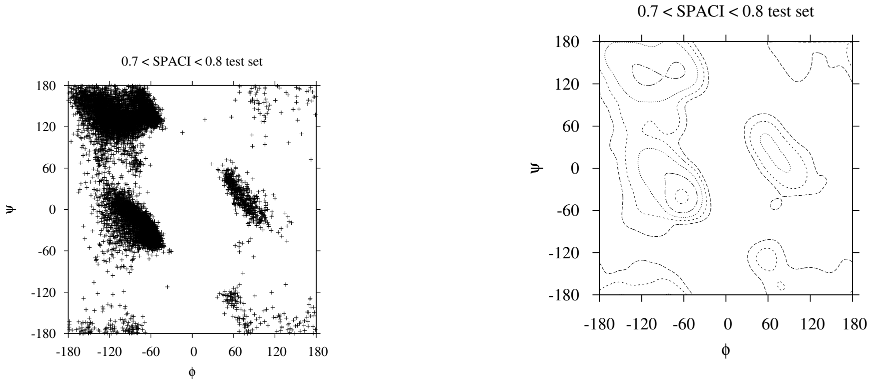


Figure S6: The Ramachandran map of the backbone dihedral angles in the test set (the data distribution). Left: backbone dihedral angles without smoothing, Right: backbone dihedral angles smoothed by Gaussians with standard deviation of 10 degrees.

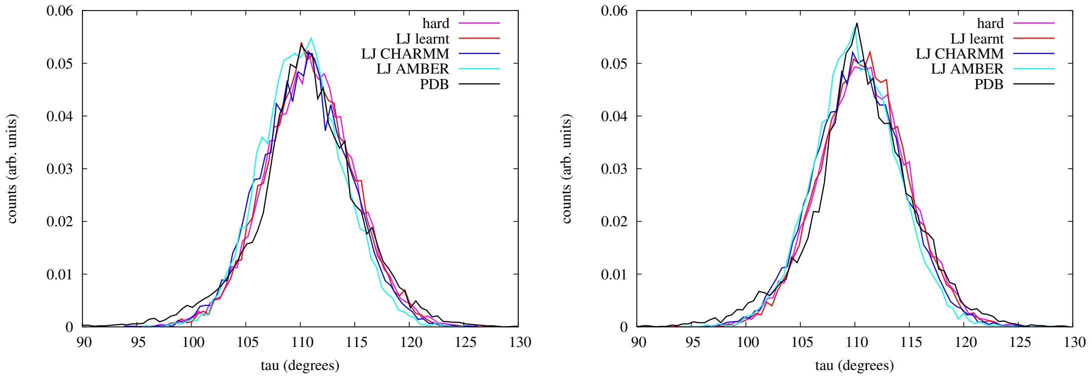


Figure S7: The distribution of the α-carbon valence angle, τ, in the data distribution and in the model distributions with the models listed in Table 1 of the main text. Left: The data distribution of the training set (black curve) with mean and standard deviation 111 ± 5 is reproduced by all model distributions of the training set (hard cutoff: 111 ± 4, purple curve; $\mathrm{LJ}_{\mathrm{learnt}}$: 111 ± 4, red curve; $\mathrm{LJ}_{\mathrm{CHARMM}}$: 111 ± 4, blue curve; $\mathrm{LJ}_{\mathrm{AMBER}}$: 110 ± 4, cyan curve. Right: The data distribution of the training set (black curve) with mean and standard deviation 111 ± 5 is reproduced by all model distributions of an independent PDB set, the ASTRAL PDB structures with SPACI 0.7–0.8 (hard cutoff: 111 ± 4, purple curve; $\mathrm{LJ}_{\mathrm{learnt}}$: 111 ± 4, red curve; $\mathrm{LJ}_{\mathrm{CHARMM}}$: 111 ± 4, blue curve; $\mathrm{LJ}_{\mathrm{AMBER}}$: 110 ± 4, cyan curve.

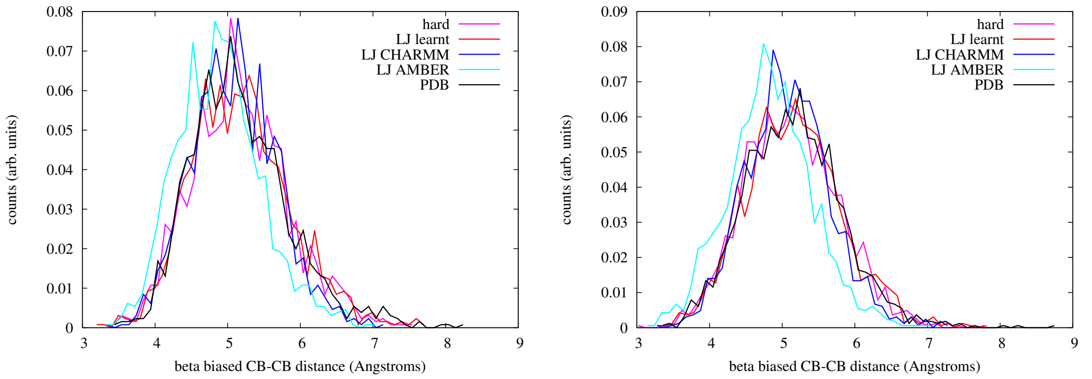


Figure S8: The distribution of β-carbon atom distances in β-sheet interactions, in the data distribution and in the model distributions with the models listed in Table 1 of the main text. Left: The data distribution of the training set (black curve) with mean and standard deviation 5.2 ± 0.7 is reproduced by all model distributions of the training set (hard cutoff: 5.2 ± 0.7, purple curve; $\mathrm{LJ}_{\mathrm{learnt}}$: 5.2 ± 0.7, red curve; $\mathrm{LJ}_{\mathrm{CHARMM}}$: 5.1 ± 0.6, blue curve; $\mathrm{LJ}_{\mathrm{AMBER}}$: 4.9 ± 0.6, cyan curve. Right: The data distribution of the training set (black curve) with mean and standard deviation 5.2 ± 0.7 is reproduced by all model distributions of an independent PDB set, the ASTRAL PDB structures with SPACI 0.7–0.8 (hard cutoff: 5.2 ± 0.6, purple curve; $\mathrm{LJ}_{\mathrm{learnt}}$: 5.2 ± 0.7, red curve; $\mathrm{LJ}_{\mathrm{CHARMM}}$: 5.1 ± 0.6, blue curve; $\mathrm{LJ}_{\mathrm{AMBER}}$: 4.9 ± 0.6, cyan curve.

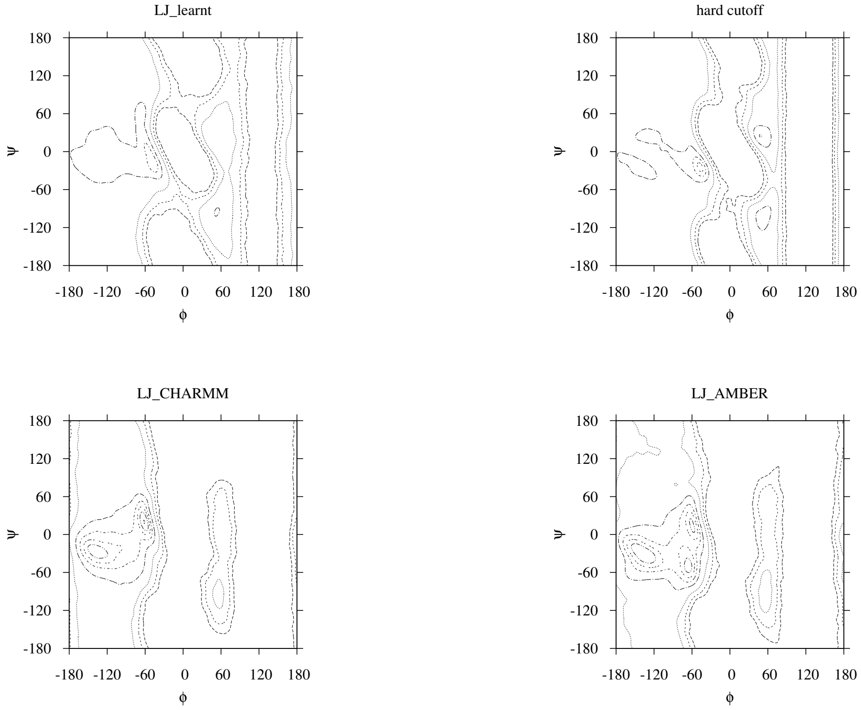


Figure S9: The distribution of the backbone dihedral angles in the MC simulations of Ala_{16} without a secondary structure bias, using the Top Left: $\text{LJ}_{\text{learnt}}$, Top Right: hard cutoff, Bottom Left: $\text{LJ}_{\text{CHARMM}}$, and Bottom Right: LJ_{AMBER} models. The distribution has been smoothed by Gaussians with standard deviation of 5 degrees.

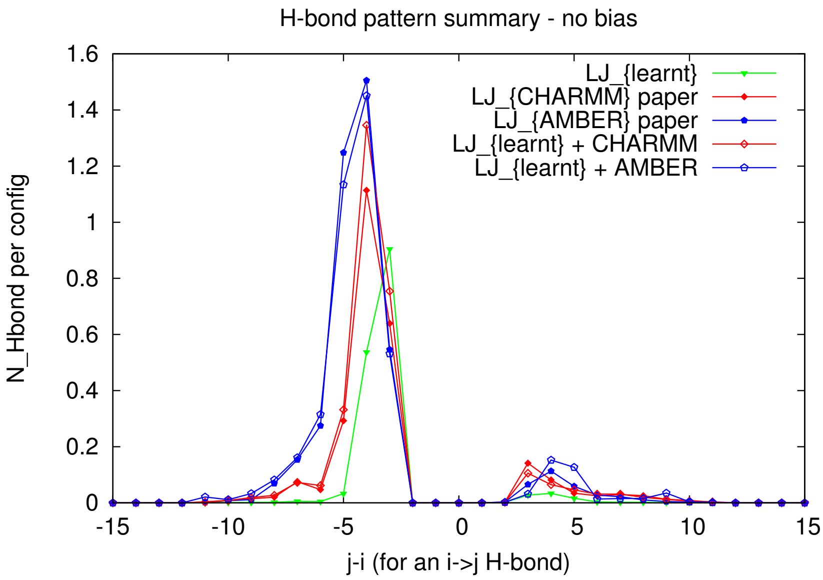


Figure S10: The H-bond pattern from MC simulations of a 16-residue polyalanine, using the $\mathrm{LJ}_{\mathrm{learnt}}$ model (green curve with filled triangles), the $\mathrm{LJ}_{\mathrm{CHARMM}}$ model (red curve with filled diamonds), the $\mathrm{LJ}_{\mathrm{AMBER}}$model (blue curve with filled pentagons), , the $\mathrm{LJ}_{\mathrm{learnt}}$ model with the vdW parameters of the $\mathrm{LJ}_{\mathrm{CHARMM}}$ model (red curve with open diamonds), and the $\mathrm{LJ}_{\mathrm{learnt}}$ model with the vdW parameters of the $\mathrm{LJ}_{\mathrm{AMBER}}$ model (blue curve with open pentagons). No secondary structure bias was applied.

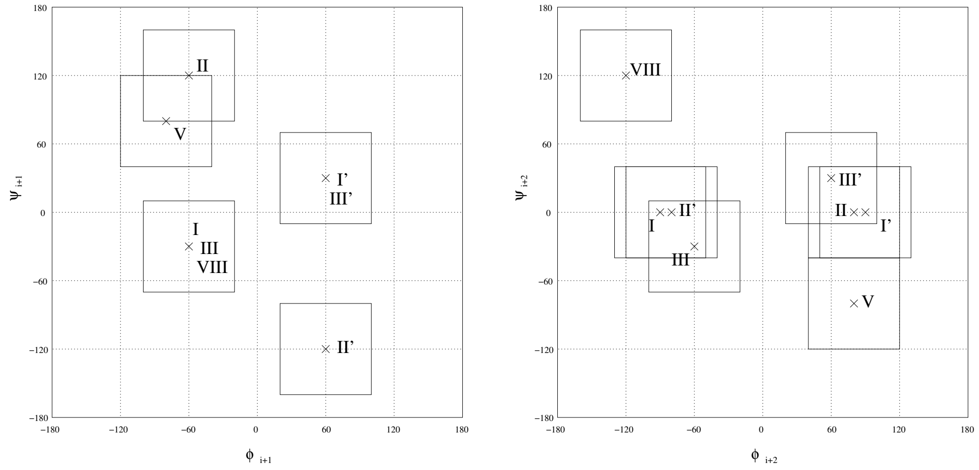


Figure S11: Turn types definitions by Venkatachalam.[S1] Left: the backbone dihedral angles of the $i+1$ residue, and Right: the backbone dihedral angles of the $i+2$ residue of the turn. Here, we allow a deviation of 40° from the ideal values, and the closest centre is taken into account when identifying the turn type. When a turn does not belong to the marked rectangles, it is type IV. Inverse turn types are marked with the prime symbol.

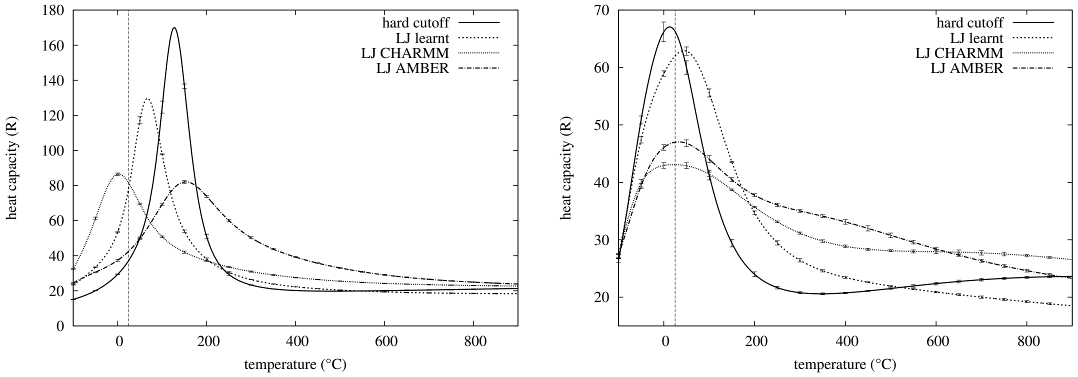


Figure S12: Heat capacity curves of Left: α-helix and Right: β-hairpin formation from NS simulations of an Ala$_{16}$ peptide with secondary structure bias, using a hard cutoff vdW potential (solid line), and a Lennard-Jones potential with learnt (dashed line) and adopted vdW parameter values from CHARMM (dotted line) and AMBER (dash-dotted line). Room temperature is marked by a vertical dashed line at 25°C.

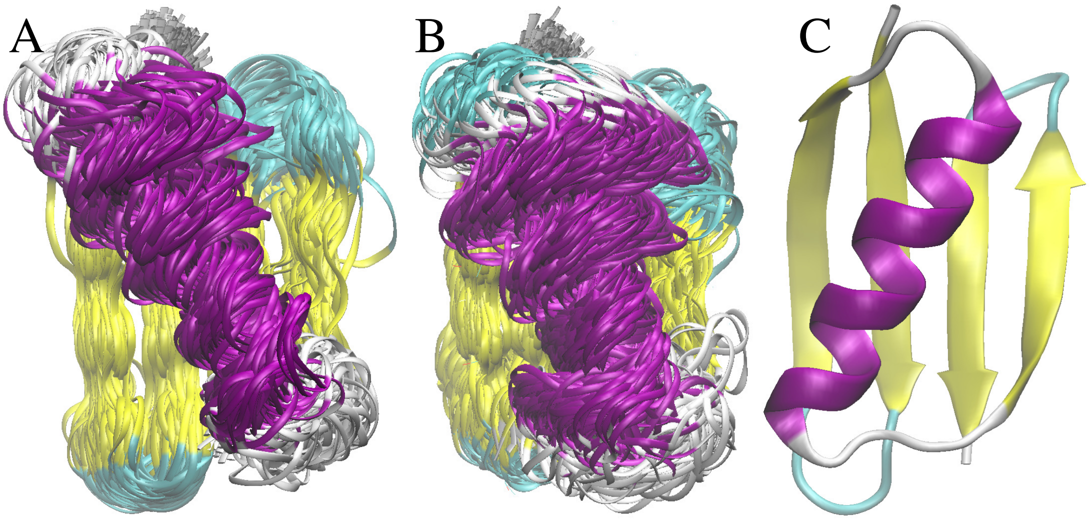


Figure S13: The ensemble of conformations at or below room temperature, in the main basin of the energy landscape, explored by Nested Sampling simulations using the protein model with A: hard cutoff vdW potential, B: Lennard-Jones type vdW potential with inferred vdW parameters. The crystal structure is shown on the right (C). Conformations obtained by using the LJ potential show a wide range of allowed helix orientation angle at room temperature, including the native angle in the crystal structure, 21.8°, while simulations using the hard cutoff potential fail to find the native helix orientation.

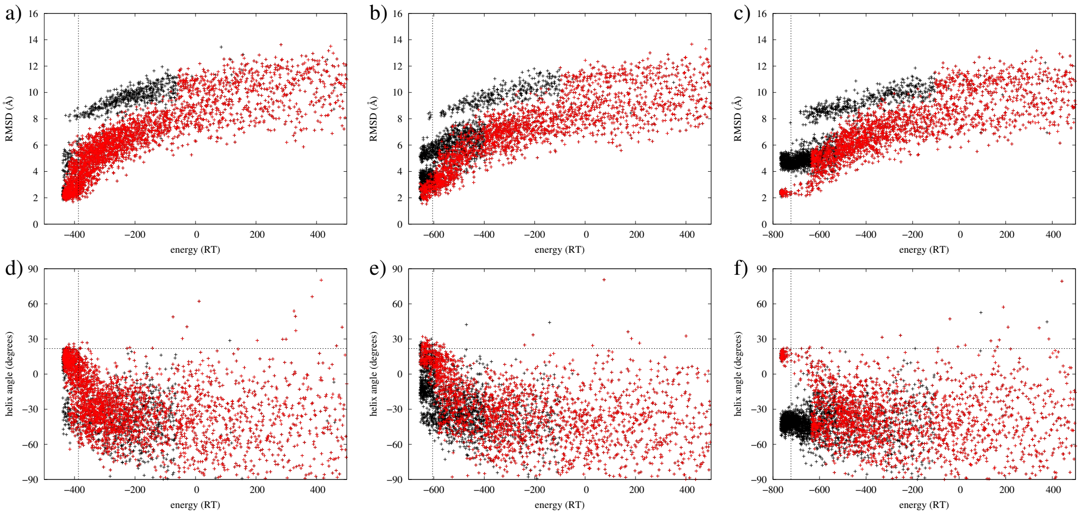


Figure S14: The backbone RMSD from the native state (a)–c)), and the angle of the helix with respect to the axis of the β-strands (d)–f)), as a function of the potential energy for all sample points (black) and sample points in the main basin of the energy landscape (red), explored by Nested Sampling simulations using the LJ type protein model with vdW parameters a) and d): inferred, b) and e): adopted from the CHARMM force field, and c) and f): adopted from the AMBER force field. The estimated energies at room temperature are marked by vertical dashed lines. The helix orientation angle in the crystal structure (21.8°) is marked by horizontal dashed lines. While all three simulations explored conformations with an RMSD of as low as 2 Å from the crystal structure, for the simulations with the LJ potential with adopted vdW parameters, the energy landscape splits above the estimated energy at room temperature, indicating a bimodal distribution of the helix orientation angle at room temperature, and for the simulation using the LJ potential with inferred vdW parameter values, the main basin split below the estimated energy at 298 K, indicating a broad unimodal distribution at room temperature.

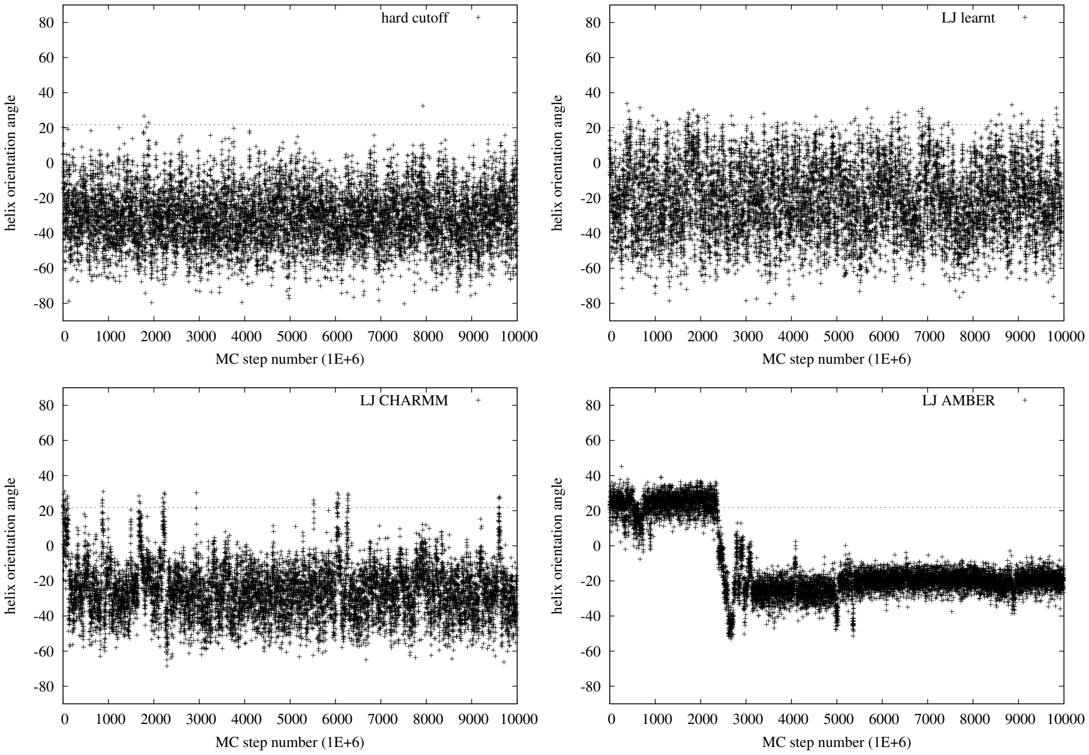


Figure S15: The trace plots of the helix orientation angle of Protein G, calculated from room temperature MC simulatilons of 10^{10} steps, using Top Left: the protein models employing the hard cutoff vdW potential, Top Right: the LJ$_{\text{learnt}}$ model, Bottom Left: the LJ$_{\text{CHARMM}}$ model and Bottom Right: the LJ$_{\text{AMBER}}$ model. The LJ$_{\text{CHARMM}}$ and the LJ$_{\text{AMBER}}$ models imply a two-state model with an energy barrier.

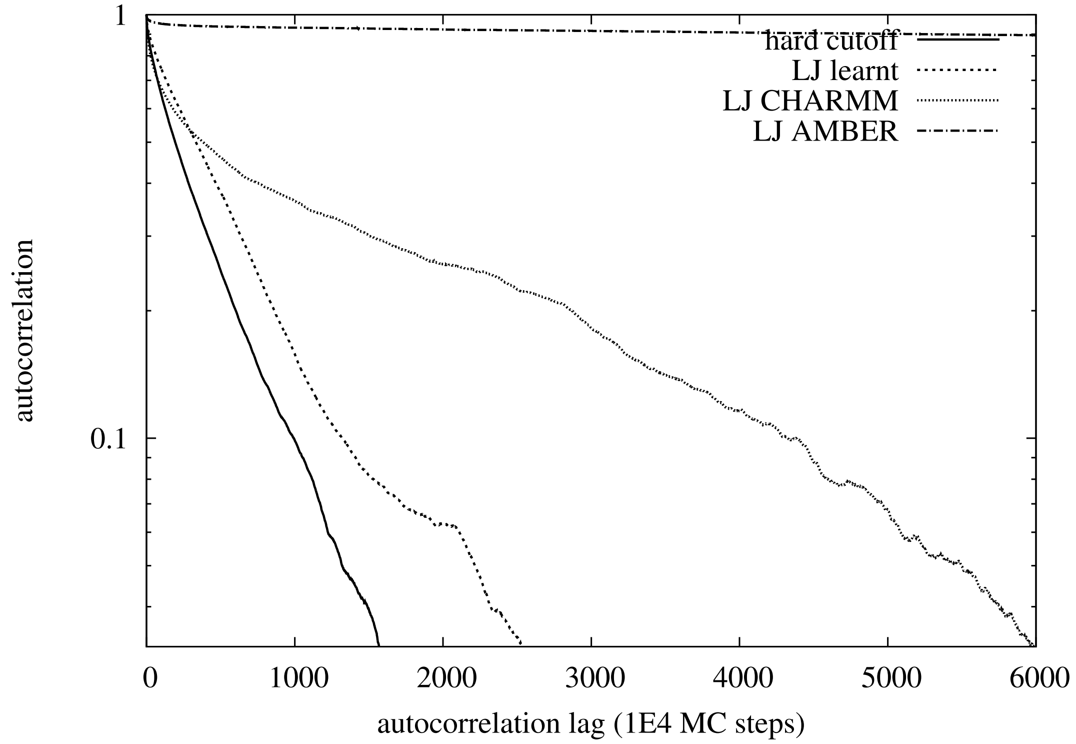


Figure S16: The autocorrelation function of the helix orientation angle of Protein G, calculated from room temperature MC simulatilons of 10^{10} steps, using the protein models employing the hard cutoff vdW potential (solid line), the $\mathrm{LJ}_{\mathrm{learnt}}$ model (dashed line), the $\mathrm{LJ}_{\mathrm{CHARMM}}$ model (dotted line) and the $\mathrm{LJ}_{\mathrm{AMBER}}$ model (dash-dotted line). The large correlation times using the $\mathrm{LJ}_{\mathrm{CHARMM}}$ and the $\mathrm{LJ}_{\mathrm{AMBER}}$ models imply a two-state model with an energy barrier.

3 Tables referenced in the Main Text

Table S1: The list of proteins used in the parameter estimation of the energy function. Listed are the ASTRAL 1.75 protein IDs with their ASTRAL classes, their SPACI scores, and the C_α RMSDs between the crystal structures and the structures mapped onto the coarse-grained CRANKITE model.

ASTRAL ID	class	SPACI	C_α RMSD (Å)
d1a6ma_	α	1.16	0.024
d1atga_	α	0.96	0.017
d1bi5a1	β	0.82	0.023
d1bi5a2	β	0.82	0.022
d1bkra_	β	1.03	0.017
d1bqka_	β	0.85	0.020
d1byia_	β	1.18	0.027
d1c0pa2	α/β	1.01	0.031
d1c1da1	α/β	0.83	0.034
d1c1da2	α/β	0.83	0.035
d1c52a_	α/β	0.93	0.023
d1c5ea_	α/β	1.09	0.027
d1c75a_	α/β	1.26	0.025
d1c9oa_	α/β	0.98	0.036
d1cc8a_	α/β	1.12	0.017
d1csei_	α/β	0.85	0.041
d1ctja_	α/β	1.06	0.023
d1cy5a_	α/β	0.90	0.019
d1d4oa_	$\alpha+\beta$	0.94	0.018
d1d4ta_	$\alpha+\beta$	1.08	0.025
d1d5ta2	$\alpha+\beta$	1.05	0.019
d1ds1a_	$\alpha+\beta$	1.02	0.021
d2a13a1	α	0.88	0.017
d2a26a1	α	1.03	0.022
d2acfa1	α	0.84	0.027
d2aeba1	α	0.81	0.022
d2akza1	α	0.87	0.044
d2akza2	α	0.87	0.045
d2b3na1	β	0.91	0.028
d2b5aa1	β	0.81	0.017
d2b69a1	β	1.02	0.029
d2b82a1	β	0.91	0.014
d2bfdb2	β	0.82	0.026
d2bhua1	β	1.05	0.032
d2bhua2	β	1.05	0.033
d2blna1	β	1.00	0.034
d2bmoa1	β	0.95	0.021
d2bmoa2	β	0.95	0.019
d2bmob1	β	0.95	0.024
d2bw4a1	β	1.27	0.022
d2bwfa1	β	0.99	0.024
d2c1ia1	α/β	0.82	0.021
d2c4ba1	α/β	0.92	0.035
d2c4ja1	α/β	0.84	0.013
d2c4ja2	α/β	0.84	0.020
d2c5aa1	α/β	0.87	0.025
d2c60a1	α/β	1.01	0.023
d2c71a1	α/β	1.12	0.018
d2cara1	α/β	1.04	0.045
d2chha1	α/β	1.16	0.036
d2ciwa2	α/β	0.97	0.019
d2cs7a1	α/β	1.06	0.022
d2d8da1	$\alpha+\beta$	1.00	0.013
d2dfba1	$\alpha+\beta$	1.04	0.036
d2dlba1	$\alpha+\beta$	0.98	0.027
d2dsxa1	$\alpha+\beta$	1.60	0.022
d3b7sa1	β	0.82	0.021
d3b7sa2	β	0.82	0.024
d3b7sa3	β	0.82	0.022
d3bbba1	β	0.87	0.024
d3bnea1	β	0.81	0.022
d3boea1	β	0.85	0.023
d3bvua1	β	1.03	0.017
d3c70a1	α/β	1.13	0.023
d3c8ya1	α/β	0.82	0.028
d3c8ya2	α/β	0.82	0.033
d3c8ya3	α/β	0.82	0.031
d3cjsb1	α/β	0.82	0.024
d3ct6a1	α/β	1.02	0.034
d3d1ka1	$\alpha+\beta$	0.90	0.017
d3d1kb1	$\alpha+\beta$	0.90	0.025
d3d32a1	$\alpha+\beta$	0.88	0.016
d7a3ha_	α	1.21	0.022

Supporting References

(S1) Venkatachalam, C. M. *Biopolymers* **1968**, *6*, 1425–1436.

Chapter 4

β-Contact Prediction for Protein Structure Prediction

CRANKITE is a weakly biased coarse-grain protein model and requires protein secondary structure and β-contacts as inputs. In the previous chapters, proteins with known tertiary structure were folded and known secondary structure and β-contacts were used. In order for CRANKITE to be used for proteins with unknown structure, for example in protein structure prediction, both protein secondary structure and β-contacts must be predicted. In this paper we develop a β-contact prediction algorithm.

As described in the Introduction, our approach uses a correlated mutation statistic, the *direct information*, which uses the correlations between columns of protein multiple sequence alignments (MSA) to predict protein contacts. Traditional correlated mutation statistics, such as mutual information, are often pairwise statistics and hence are unable to distinguish between protein contacts and transitive correlations. The direct information, however, is more global in nature and does not suffer from this problem. In order to calculate the statistic, an optimization procedure is required, and many approaches have been used. In this work, however, we use the contrastive divergence approach described in Chapter 3.

In contrast to many correlated mutation analysis studies, we do not focus on a small group of protein families with large, high quality alignments (for which the structure of at least one member of the family almost always exists, and hence template-based methods could be used). Instead we take a standard dataset of 916 proteins commonly used for benchmarking β-contact prediction algorithms. The proteins in this dataset have varying sizes of MSAs.

We show that the correlated mutation statistic contains information, even for proteins with small sequence alignments, but by itself is unlikely to provide enough information to generate the tertiary structure of the protein. We show that by coupling the direct information to an empirical Bayes β-sheet model we can make use of the correlated mutation information, significantly improving the accuracy of the model. The strong constraints and prior knowledge associated with β-contacts nicely complement the often noisy direct information.

We compare our algorithm to state-of-the-art β-contact prediction algorithms producing competitive results both at the residue and strand level. The only information our algorithm requires from the MSA is the direct information and we thus show that the direct information is a very powerful statistic, as other approaches require the entire MSA to be input, together with extra information such as residue solvent accessibility and sequence position of the helices of the proteins, although their results are no better than ours.

Following previous β-contact work, the algorithm is initially benchmarked assuming protein secondary structure is known. However, in the supporting information, we also test the algorithm with the CASP 2010 dataset using *predicted* secondary structure. Finally, linking this work to Chapters 2 and 3, we use nested sampling CRANKITE simulations, with predicted β-contacts as inputs, to successfully fold two of these CASP proteins.

4.1 Contribution

The initial idea was formulated by Burkoff and Wild. Burkoff, under the supervision of both Wild and Várnai, developed and implemented the model and benchmarked the dataset. The initial draft of the paper was written by Burkoff, with all authors contributing to the revision of the paper before submission. Burkoff also drafted a response to referees with input from Wild and Várnai during the peer review process and revised the paper before publication.

The article was originally published by Oxford University Press: Burkoff, N. S. , C. Várnai and D. L. Wild. Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. *Bioinformatics* (2013) 29(5)580–587.

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 5 2013, pages 580–587
doi:10.1093/bioinformatics/btt005

Structural bioinformatics

Advance Access publication January 10, 2013

Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure

Nikolas S. Burkoff, Csilla Várnai and David L. Wild*

Systems Biology Centre, Senate House, University of Warwick, Coventry, CV4 7AL, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The problem of *ab initio* protein folding is one of the most difficult in modern computational biology. The prediction of residue contacts within a protein provides a more tractable immediate step. Recently introduced maximum entropy-based correlated mutation measures (CMMs), such as direct information, have been successful in predicting residue contacts. However, most correlated mutation studies focus on proteins that have large good-quality multiple sequence alignments (MSA) because the power of correlated mutation analysis falls as the size of the MSA decreases. However, even with small autogenerated MSAs, maximum entropy-based CMMs contain information. To make use of this information, in this article, we focus not on general residue contacts but contacts between residues in β-sheets. The strong constraints and prior knowledge associated with β-contacts are ideally suited for prediction using a method that incorporates an often noisy CMM.

Results: Using contrastive divergence, a statistical machine learning technique, we have calculated a maximum entropy-based CMM. We have integrated this measure with a new probabilistic model for β-contact prediction, which is used to predict both residue- and strand-level contacts. Using our model on a standard non-redundant dataset, we significantly outperform a 2D recurrent neural network architecture, achieving a 5% improvement in true positives at the 5% false-positive rate at the residue level. At the strand level, our approach is competitive with the state-of-the-art single methods achieving precision of 61.0% and recall of 55.4%, while not requiring residue solvent accessibility as an input.

Availability: http://www2.warwick.ac.uk/fac/sci/systemsbiology/research/software/

Contact: D.L.Wild@warwick.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 27, 2012; revised on December 17, 2012; accepted on January 2, 2013

1 INTRODUCTION

The problem of *ab initio* protein folding is one of the most difficult in modern computational biology. The prediction of residue contacts within a protein provides a more tractable immediate step, and these contacts can be used as a guide to generate the tertiary structure of the protein.

Correlated mutation (CM) methods, first pioneered by Valencia and colleagues (Gobel *et al*., 1994), take a multiple sequence alignment (MSA) profile of evolutionarily related proteins and attempt to predict residues that have co-evolved. If residues have co-evolved, this may imply proximity in the native structure. For example, if a small residue increases in size by mutating, a proximal residue may have to reduce in size to retain the viability of the fold.

Many CM methods have been developed using Pearson correlation coefficients (Gobel *et al*., 1994), adaptions of Mutual Information (Dunn *et al*., 2008; Lee and Kim, 2009), perturbation methods (Dekker *et al*., 2004) and Dynamic Bayesian networks (Burger and van Nimwegen, 2010).

A recently developed correlated mutation measure (CMM), the *direct information* (Morcos *et al*., 2011; Weigt *et al*., 2008), is a global measure that is derived from modelling the entire MSA, specifically defining the probability of each sequence being a member of the MSA. This distribution shares the same low-order moments as the MSA, and the maximum entropy principle (Jaynes, 2007) is used to fully specify the distribution. Marks *et al*. (2011), Sułkowska *et al*. (2012) and Hopf *et al*. (2012) have used this measure to successfully aid the folding of a diverse range of proteins. However, like the majority of CM studies, these authors focused on a small number of proteins for which there is a large high-quality MSA because all CMMs suffer as the size of the MSA decreases (Olmea and Valencia, 1997). A key distinction of this work is that we focus on a wide selection of proteins that have a variety of sizes of MSAs. We also automate the generation of MSAs and do not rely on a large high-quality MSA being available.

In an attempt to improve the power of CM methods, the Dynamic Bayesian network of Burger and van Nimwegen (2010) incorporates primary-sequence distance into an informative prior for the model. The incorporation of this knowledge substantially improves the results. Inspired by this, we have chosen to predict the lateral pairs of residues in interacting β-strands, β-contacts, using a CMM. β-contacts are associated with strong constraints, for example, sequential pairs of residues form β-contacts and residues can only be in β-contact with up to two other residues. These constraints mean β-contacts are ideally suited for prediction using a CMM—the noise associated with the CMM is compensated for by incorporating the strong β-contact constraints.

The prediction of β-contacts can be used to aid tertiary structure prediction (Podtelezhnikov and Wild, 2009; Ruczinski *et al*., 2002), explore energy landscapes (Burkoff *et al*., 2012), in designing proteins (Kortemme *et al*., 1998; Smith and Regan, 1995) and understanding protein folding pathways (Mandel-Gutfreund *et al*., 2001; Merkel and Regan, 2000).

*To whom correspondence should be addressed.

We highlight BetaPro, the work of Cheng and Baldi (2005), which uses a three-stage method to predict β-topologies and was the first method to take into account the global nature of β-topologies. Firstly, a 2D recurrent neural network is used to generate a residue-level pairing map. Secondly, a dynamic programming algorithm is applied to this map to derive strand-level pseudo binding energies and finally, a graph matching algorithm is used to predict strand contacts.

There are a variety of other existing methods for β-contact prediction. They include the use of statistical potentials (Hubbard and Park, 1994), information theoretic approaches (Steward and Thornton, 2002), integer linear optimization (Rajgaria *et al.*, 2010), hybrid neural network-probabilistic models (Aydin *et al.*, 2011) and Markov logic networks (MLNs; Lippi and Frasconi, 2009).

In this article, we have developed a global probabilistic model for β-contact prediction, inspired by the secondary structure models of Schmidler (2002), which can be used to predict both residue- and strand-level interactions. We have integrated this model with a CMM, similar in nature to direct information, and using this model on a standard dataset, significantly outperform the recurrent neural network of BetaPro and are competitive with the best single methods currently available. Unlike these methods, our approach does not require additional information such as residue solvent accessibility to be entered as an input to the model. In common with other methods, we assume the native secondary structure is known. However, our framework can be easily extended to predict both secondary structure and β-contacts simultaneously, and this is the focus of our current work.

2 METHODS

2.1 Data set

In this work, we use the set of 916 proteins from Cheng and Baldi (2005) (CB916). The proteins share no >15–20% sequence identity, and the set consists of 187 516 residues, of which 48 996 are strand residues, which are involved in 31 638 β-contacts.

Most CM analysis procedures focus primarily on proteins for which there is a large good-quality MSA, often a large PFAM alignment (Sonnhammer *et al.*, 1997). We wanted to develop a method that will take advantage of this information where it exists, and yet is applicable even if the CM analysis is not useful, or indeed there is no MSA, which can be the case for newly sequenced proteins, such as those selected as targets in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) community-wide experiment. Therefore, our method of generating MSAs is extremely general.

We generate MSAs following a similar method to Saqi *et al.* (1999). For each sequence, we run PSI-Blast (Altschul *et al.*, 1997) for two iterations (Evalue = 0.005) against the Non-Redundant database, keeping all sequences that share at least 30% identity to the profile constructed after the first PSI-Blast iteration, similar to the procedure recommended in Ashkenazy *et al.* (2009). We then perform a global–local alignment using GLsearch (Pearson, 2000) to trim the sequences PSI-Blast found. We then use CD-Hit (Li *et al.*, 2001; Li and Godzik, 2006) to cluster the trimmed sequences at the 98% threshold and use Muscle (Edgar, 2004) (maxiters = 2) to generate MSAs. Finally, we removed columns of the MSA that were gaps in our target sequence and any row that contained >33% gaps. There is an enormous variation in the number of sequences in the alignments: six proteins have no homologues, one-fifth have <100 homologues and 7% have >2000 (see Fig. 1).

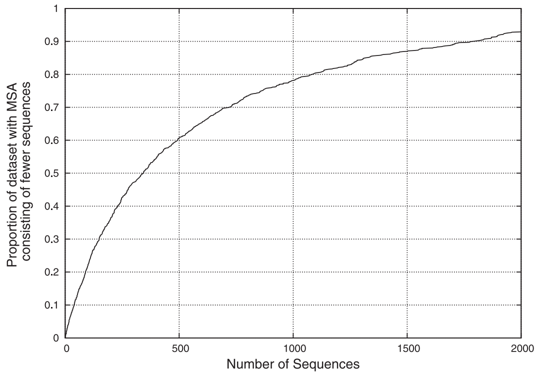


Fig. 1. The number of sequences in the 916 MSAs varies enormously. For example, 60% of the MSA have <500 sequences

2.2 Maximum entropy-based CM measure

CMMs based on maximum entropy modelling (also called Direct Coupling Analysis) (Marks *et al.*, 2011; Morcos *et al.*, 2011; Weigt *et al.*, 2008) aim to distinguish between *direct* and *indirect* correlations. Direct correlations arise owing to proximity in the native structure of the protein and are of primary interest in contact prediction; indirect correlations are caused by other reasons, such as the fact that correlations are transitive, and are the cause of the poor performance of many CMMs.

The idea is to model the entire family of evolutionarily related proteins, assigning probability mass over all possible (fixed-length) sequences, including those that have not been observed. From this global model, measures can be developed to model the strength of the direct correlations between pairs of residues. This idea is formalized below.

Given an MSA containing M sequences for a protein of length N, we define $f_i(A_i)$ as the observed frequency of residue A_i occurring in position i of the MSA and $f_{ij}(A_i, A_j)$ as the observed frequency of both residue A_i occurring in position i and residue A_j occurring in position j of the MSA. Given any sequence $\mathbf{A} = A_1, A_2, \ldots, A_N$, we model the probability of it occurring in the MSA by a distribution $P(\mathbf{A}) = P(A_1, A_2, \ldots, A_N)$ However, there are q^N possible different sequences (where q is the size of the alphabet of amino acids) and only $M \ll q^N$ sequences in the MSA. The sparsity of the data and the number of sequences imply that it is impractical for detailed use. However, we would like our model to match the empirical low-order moments given by the MSA. Specifically we would like

$$P_i(A_i) = f_i(A_i) \quad \text{and} \quad P_{ij}(A_i, A_j) = f_{ij}(A_i, A_j)$$

where $P_i(.)$ is the marginal distribution for position i and $P_{ij}(.,.)$ is the (joint) marginal distribution (We have not added pseudo-counts or weighted sequences) marginal distribution of positions i and j (we have not added pseudo-counts or weighted sequences).

Among the valid distributions P satisfying these constraints, using the maximum entropy principle (Jaynes, 2007), we favour $\mathfrak{P}$, the distribution that has maximum entropy, S:

$$\mathfrak{P} = \text{argmax}_P[S(P)] \equiv \text{argmax}_P\left\{-\sum_{\mathbf{A}} P(\mathbf{A})\log[P(\mathbf{A})]\right\}$$

and solving this optimization problem using Lagrange multipliers leads to the distribution

$$\mathfrak{P}(A_1, \ldots, A_N) \propto \exp\left[-\sum_{1 \le i \le j \le N} e_{ij}(A_i, A_j) + \sum_{1 \le i \le N} h_i(A_i)\right]$$

for some pair-interaction energies $e_{ij}(A_i, A_j)$ and local fields $h_i(A_i)$ (Weigt *et al.*, 2008). See the Supplementary Data for further details.

The maximum entropy distribution can be viewed as a Potts model on an underlying complete graph, where the nodes represent the residue positions, the ‘spins’ correspond to the amino acid types and the edges describe the pairwise interactions, whose strengths are described by the pairwise interaction energies e_{ij}. A related model for protein families, using Markov random fields (Balakrishnan *et al.*, 2011), can also be viewed as a Potts model. However, instead of the underlying graph being complete, an optimal subgraph is chosen that aims to fully explain the correlations and conditional independencies within the underlying protein family.

To generate the maximum entropy distribution $\mathfrak{P}$, we use a statistical machine learning technique, contrastive divergence (Hinton, 2002). This work represents the first application of this approach to the modelling of protein MSAs. For a given set of $e_{ij}(A_i, A_j)$ and $h_i(A_i)$, we use contrastive divergence to approximate the marginal distributions $P_i(.)$ and $P_{ij}(.,.)$ and use gradient descent to update e_{ij} and h_i. We iterate this procedure to convergence. For a protein of 75 residues, the procedure takes ~10 minutes on a single core of an Intel Core i7 processor, and for a protein of 350 residues, the procedure takes ~2.5 h. Further details are found in the Supplementary Data.

Once we have calculated the distribution $\mathfrak{P}$, we define our CMM, $\mathcal{D}$. For each pair of residues (i, j), we define $\mathcal{D}(i,j)$ as follows:

$$\mathcal{D}(i,j) = \sum_{A_i, A_j} \mathfrak{P}^D_{ij}(A_i, A_j) \log \frac{\mathfrak{P}^D_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)}$$

where

$$\mathfrak{P}^D_{ij}(A_i, A_j) \propto f_i(A_i) f_j(A_j) \exp[-e_{ij}(A_i, A_j)].$$

This is a modified version of the *Direct Information* previously used to predict protein contacts (Marks *et al.*, 2011; Weigt *et al.*, 2008). The Direct Information measure itself was tried but produced slightly poorer results than $\mathcal{D}$. See the Supplementary Data for more details.

To show the power of $\mathcal{D}$, for each protein in the dataset, we took the top $N/2$ ranked $\mathcal{D}(i,j)$, where N is the length of the protein (we remove those for which $|i-j| \leq 4$ from the analysis) and calculated the contact ratio: the proportion of these pairs of residues whose C_α distance is ≤ 8 Å. The contact ratio versus $\log(M)$ is shown in Figure 2 (Top). Figure 2 (Bottom) shows the average C_α distance of these $N/2$ predicted contacts. These figures show that there is a lot of information contained within $\mathcal{D}$, especially as M increases.

However, using randomly chosen contacts of known structures, it has been shown that one needs around a quarter to two-fifths of contacts to be able to successfully regenerate the native structure (Duarte *et al.*, 2010; Sathyapriya *et al.*, 2009; Vendruscolo *et al.*, 1997). Marks *et al.* (2011) and Hopf *et al.* (2012) have shown that if a protein has a large number of sequences in its MSA, then maximum entropy-based CM analysis, together with predicted secondary structure is enough to successfully reconstruct the tertiary structure of the protein. In these articles, the authors take the highest-ranked correlated pairs of residues to be incorporated into distance constraints used to generate initial all-atom conformations of the protein. Simulated annealing, relaxing these distance constraints throughout the simulation, is then used to generate final three-dimensional structures.

However, as shown by Figures 1 and 2, a large number of proteins have only a small MSA and CMMs by themselves are unlikely to be able to provide a large enough number of contacts to successfully fold the protein. For example, Marks *et al.* (2011) restrict their attention to proteins whose MSA has at least 1000 sequences, and usually significantly more. Nevertheless, even an alignment with $M = e^6 \approx 400$ sequences produces an average contact ratio of ~0.15, which still contains lots of information (for an average protein, the contact ratio for randomly chosen contacts is ~0.03). In contrast to these other studies, we investigate whether one can make use of this evolutionary information. We propose to use $\mathcal{D}$ to improve the prediction of β-contacts, for which there is a large amount of structural knowledge, which can be incorporated as prior beliefs within a Bayesian statistical framework. The following sections describe the new β-strand Bayesian model we have developed and how we couple $\mathcal{D}(i,j)$ to it.

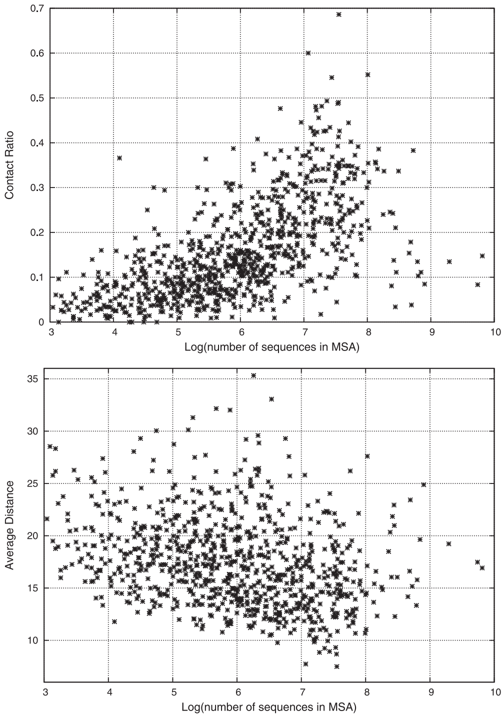


Fig. 2. Top: The proportion of the top $N/2$ ranked $\mathcal{D}(i,j)$ in contact (contact ratio) versus the number of sequences in the MSA for each protein in the CB916 dataset. A contact is defined as the C_α distance being ≤ 8 Å. Bottom: The average C_α distance of the top $N/2$ ranked $\mathcal{D}(i,j)$. Two outliers [at (3.4, 37.9) and (4.5, 42.1)] are not shown

2.3 β-Topology model

Given a primary sequence $\mathbf{R} = \{R_1, R_2, \ldots, R_N\}$ and its secondary structure $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$, where S_i is the secondary structure of residue i, residues R_i and R_j are defined to be a β-contact if they are a lateral pair within two interacting β-strands. For example, in Figure 3, residues 6 and 53 are a parallel β-contact and residues 44 and 53 are an antiparallel β-contact. We define $\mathcal{I}$ to be the set of β-contacts. Specifically $(i, j, 1) \in \mathcal{I}$ if residues R_i and R_j are a parallel β-contact, $(i, j, -1) \in \mathcal{I}$ if residues R_i and R_j are an antiparallel β-contact and $(i, j, 0) \in \mathcal{I}$ if either residue R_i or R_j is an isolated β-bridge. We say $(i, j) \in \mathcal{I}$ if $(i, j, 1)$, $(i, j, 0)$ or $(i, j, -1) \in \mathcal{I}$.

The general framework we are using (from Schmidler, 2002) allows inference for $\mathbf{S}$ and $\mathcal{I}$ given $\mathbf{R}$. Following the Bayesian method, we require a prior $\mathbb{P}(\mathbf{S}, \mathcal{I}) = \mathbb{P}(\mathcal{I}|\mathbf{S})\mathbb{P}(\mathbf{S})$ and a likelihood $\mathbb{P}(\mathbf{R}|\mathbf{S}, \mathcal{I})$. Using Bayes’ theorem, these yield the posterior of interest $\mathbb{P}(\mathbf{S}, \mathcal{I}|\mathbf{R}) \propto \mathbb{P}(\mathbf{R}|\mathbf{S}, \mathcal{I})\mathbb{P}(\mathbf{S}, \mathcal{I})$.

In this work, we assume the secondary structure is fixed. Specifically $\mathbb{P}(\mathbf{S}) = 1$, if $\mathbf{S}$ is the secondary structure assignment given by DSSP (Kabsch and Sander, 1983)—we map residues labelled E and B to E, strand residues, and all other labels to C, non-strand residues. For clarity we suppress the dependency on $\mathbf{S}$, i.e. $\mathbb{P}(\mathcal{I}|\mathbf{S}) = \mathbb{P}(\mathcal{I})$. A focus of our current work is to extend the model to allow joint inference for $\mathbf{S}$ and $\mathcal{I}$.

Definitions: Viewing $\mathcal{I}$ as a collection of individual residue contacts does not easily allow the incorporation of the structure of β-contacts into a model; therefore, we model $\mathcal{I}$ as a set of interacting strand segments, following (Chu *et al.*, 2006). The set of residue contacts in $\mathcal{I}$ can be uniquely determined by specifying which strand segments interact and for each pair of interacting strands specifying their direction, alignment and position of any bulges. We formalize these terms below.

The strand residues of a protein can be represented as a set of distinct strand segments (For some proteins, DSSP defines two separate strand segments immediately adjacent in sequence. For example 'EEEB'. For a fair comparison with BetaPro we define a strand segment as a contiguous block of strand residues. However, this is not necessary for our model). For example, Figure 3 shows 4 strand segments (E_1,E_2,E_3,E_4). In this protein, there is a single sheet, and in this simple case, the strand interactions can be described by a permutation ϕ of the set of strand segments. Specifically $\phi(1, 2, \ldots, m) = [\phi(1), \phi(2), \ldots, \phi(m)]$ and implies segment $E_{\phi(r)}$ and $E_{\phi(r+1)}$ interact for $r = 1, 2, \ldots, m-1$. In Figure 3, $\phi(1, 2, 3, 4) = (3, 4, 1, 2)$. In more complicated cases, the sheet structure cannot be described by a permutation. For example, if there is more than one sheet, if strands are involved in more than two interactions or if there is a cycle (for example in β-barrels, where every strand interacts with two partners).

Following the terminology in (Ruczinski *et al.*, 2002), we say there is a jump between segments E_r and E_{r+1} if E_r and E_{r+1} are not interacting. In Figure 3, there is a jump between segments E_2 and E_3 and no other jumps. We define the jump pattern J as the set of r for which E_r and E_{r+1} are not interacting; in Figure 3, the jump pattern $J = \{2\}$. See Figure 4(a–d) for further examples of ϕ and J.

We introduce d_{rs} to describe the direction of interaction, specifically $d_{rs} = 1$ if interacting segments E_r and E_s are a parallel strand interaction and $d_{rs} = -1$ if the segments are antiparallel. In Figure 3, $d_{34} = d_{12} = -1$ and $d_{14} = 1$. If either E_r or E_s is an isolated β-bridge, then $d_{rs} = 0$.

The variable a_{rs} is used to define the shift between strands. For parallel interactions, a_{rs} describes the shift between the final residues of both strands. For example, in Figure 3, $a_{14} = 0$ because $(8, 55) \in \mathcal{I}$. If E_1 was shifted up by one residue, so that $(8, 54) \in \mathcal{I}$, then a_{14} would equal $+1$. Conversely, if E_1 was shifted down by two residues, so that residue $(6, 55) \in \mathcal{I}$, then a_{14} would equal -2. For antiparallel interactions, a_{rs} describes the shift between the end of the strand earlier in the sequence and the beginning of its interacting partner (i.e. between residues 8 and 13 for a_{12} in Fig. 3).

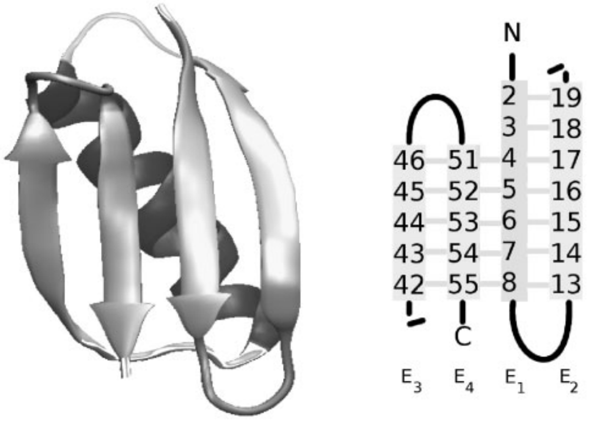


Fig. 3. Left: The structure of protein G (1PGA). Right: The β-topology for protein G. The numbers are the positions in the sequence of the strand residues and the horizontal lines are β-contacts. Residues 6 and 53 are a parallel β-contact and residues 53 and 44 are an antiparallel β-contact. Hence $(6, 53, 1)$ and $(44, 53, -1) \in \mathcal{I}$. The 4-strand segments E_1,E_2,E_3,E_4 are ordered from left to right in the sheet E_3,E_4,E_1,E_2, and hence the permutation ϕ, which permutes (1, 2, 3, 4) to (3, 4, 1, 2) describes the set of interactions. There is a jump between segments E_2 and E_3. Segments E_1 and E_2 are an antiparallel interaction and hence $d_{12} = -1$, and segments E_1 and E_4 are parallel and so $d_{14} = 1$

Restricting the number of bulges to at most one per β-strand interaction (which is the case in 98.6% of cases), we can define $b_{rs} = 0$ if there is no bulge or $b_{rs} = k$ if residue k is the β-bulge. There are no bulges in the sheet shown in Figure 3. Figure 4 shows the values of $\{d_{rs}, a_{rs}, b_{rs}\}$ for different interacting segments.

Prior for $\mathcal{I}$, $\mathbb{P}(\mathcal{I})$*:* There is a huge amount of structure in β-topologies and the challenge for a Bayesian statistician is to try and capture this while being able to efficiently calculate posterior probabilities and not overfitting the model. Rather than aim for the most probable β-topology, we calculate $\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R})$, producing a *probability contact map*, analogous to the output from BetaPro's Neural Network. Unlike other statistical models (Aydin *et al.*, 2011), we do not take the output from BetaPro's Neural Network as an input to our model.

We take advantage of the framework of Bayesian inference, which allows us to exercise our scientific judgement and experience concerning parameters that we expect to be of particular importance, and by specifying how these are plausibly related.

We model the interacting β-strands as a single sheet defined by a permutation ϕ, as described above. Although our approach does not model more than one sheet per protein, we can predict multiple sheets (see Figure 8). More complicated models involving partitioning the segments into different sheets were tried, but these did not improve the results. We only allow a single bulge per strand interaction.

Our prior is defined as

$$\mathbb{P}(\mathcal{I}) = \mathbb{P}(\phi) \prod_{r,s} \mathbb{P}(d_{rs}) \mathbb{P}(a_{rs}|d_{rs}) \mathbb{P}(b_{rs}|a_{rs}, d_{rs})$$

where the product is over all segments E_r and E_s that are interacting, given permutation ϕ, and we have suppressed the dependence of everything on the secondary structure $\mathbf{S}$. The set $\{\phi, d_{rs}, a_{rs}, b_{rs}\}$ gives a unique set of residue contacts $(i,j) \in \mathcal{I}$, and if $\mathcal{I}$ cannot be described by a set $\{\phi, d_{rs}, a_{rs}, b_{rs}\}$, then $\mathbb{P}(\mathcal{I}) = 0$. We define the distance τ_{rs} as the number of residues between segments E_r and E_s. For example, in Figure 3, $\tau_{12} = 4$, and we define l_r as the number of residues in segment E_r.

- $\mathbb{P}(\phi)$: The probability of a specific permutation depends on all the distances τ_{rs} and the lengths of all the strands l_r. However, incorporating all this information leads to an exponential number of parameters. In the dataset, 50% of interacting strands are adjacent in sequence (and 42% of adjacent strands are interacting), so one of the most important things we would like the distribution to capture is whether adjacent strands are in contact. For these reasons, in our

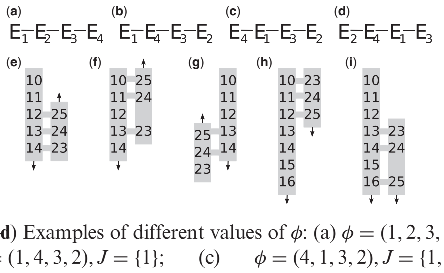


Fig. 4. **(a–d)** Examples of different values of ϕ: (a) $\phi = (1, 2, 3, 4), J = \{\}$; (b) $\phi = (1, 4, 3, 2), J = \{1\}$; (c) $\phi = (4, 1, 3, 2), J = \{1, 3\}$; (d) $\phi = (2, 4, 1, 3), J = \{1, 2, 3\}$. **(e–i)** Examples of different $\{d_{rs}, a_{rs}, b_{rs}\}$: (e) $\{-1, 0, 0\}$; (f) $\{-1, -1, 12\}$; (g) $\{-1, 1, 0\}$; (h) $\{1, -4, 0\}$; (i) $\{1, 0, 15\}$

model, all ϕ that share the same jump pattern J are equally likely, and the probability $\mathbb{P}(r \in J|\tau_{rr+1}, l_r, l_{r+1})$ is independent for each r. $\mathbb{P}(r \in J|\tau_{rr+1}, l_r, l_{r+1})$ is taken from the training set by counting occurrences. For small l and τ, we take values directly from the training set, and for larger l and τ, owing to sparsity of data, we collapse the data into a small number of bins. We added a pseudo-count to smooth the data from the training dataset.

- If either l_r or $l_s = 1$, then $\mathbb{P}(d_{rs} = 0) = 1$, otherwise $\mathbb{P}(d_{rs} = 1|\tau_{rs})$ is a piecewise linear function of τ_{rs}, fitted from the training set.
- $\mathbb{P}(a_{rs}|d_{rs})$: There is an inherent asymmetry in our definition of a_{rs}; in the case of parallel strands, we are measuring the shift from a perfect alignment of the ends of the segments, not the beginnings. In proteins, it is found that the shift measured from a perfect alignment of *at least one* end of the segments is small. Compare a_{rs} for Figures 4(h–i). Previous work has not taken this into account (Chu *et al.*, 2006), which leads to a drop in performance. Therefore, we model $\mathbb{P}(a_{rs}|d_{rs})$ as a mixture (equally weighted) of the distributions $\mathbb{P}_{d_{rs}}(a_{rs})$ and $\mathbb{P}_{d_{rs}}(\hat{a}_{rs})$, where $\hat{a}_{rs}$ is the shift required from aligning the beginnings of the segments to get the same residue contacts as a shift of a_{rs} produces from aligning the ends of the strands. These distributions are taken from the training set. An analogous procedure is followed for the antiparallel case.
- $\mathbb{P}(b_{rs} \neq 0|d_{rs}, a_{rs}) = \mathbb{P}(b_{rs} \neq 0)$ is taken from the training set, and if there is a bulge, there is a uniform probability over all residues involved in the interaction that they are a bulge (hence the dependence on d_{rs} and a_{rs}—to know which residues can be the bulge).

Likelihood: $\mathbb{P}(\mathbf{R}|\mathcal{I})$

$$\mathbb{P}(\mathbf{R}|\mathcal{I}) \propto |\mathcal{I}|^{(u|E|-1)} \exp(-v|\mathcal{I}|) \prod_{(i,j,d_{ij}) \in \mathcal{I}} \mathbb{L}(R_i, R_j|d_{ij})$$

where the joint likelihood $\mathbb{L}(R_i, R_j|d_{ij})$ is approximated from the limited training set by the product of the conditionals, $\mathbb{P}(R_i|R_j, d_{ij})$ and $\mathbb{P}(R_j|R_i, d_{ij})$, where $\mathbb{P}(.|R_j, d_{ij})$ is the distribution of amino acids in contact with the residue type of R_j in the direction of d_{ij}. $|\mathcal{I}|$ is the number of contacts and $|E|$ is the number of β-residues. The distributions $\mathbb{P}(.|R_j, d_{ij})$ are taken from the training set, and u and v are constants to be determined.

We have chosen this likelihood because of its simplicity. More complicated dependencies, such as letting R_i depend on $R_{j\pm1}$, were tried, but did not noticeably improve the results. We include a gamma distribution on the number of contacts into the likelihood because, without this term, the likelihood is a product of $2|\mathcal{I}|$ numbers smaller than one, and so actively penalises against contacts. We include $|E|$ so that the mean and variance of the gamma distribution depend on the number of β-residues, which allows the model to control the total number of contacts. This is important as $|\mathcal{I}|$ and $|E|$ are strongly correlated. The constants u and v were fitted using an empirical Bayes approach, and set to 18 and 12, respectively. See Supplementary Data for more details.

2.4 Integrating CM measure with the β-topology model

In this work, we perform inference on both the posterior distribution $\mathcal{P}_1(\mathcal{I}|\mathbf{R}) \propto \mathbb{P}(\mathcal{I})\mathbb{P}(\mathbf{R}|\mathcal{I})$ and, by adapting the concept of a 'product of experts' (Hinton, 1999, 2002), on a distribution that couples $\mathcal{D}(i,j)$ to the β-topology model. A product of experts allows different probabilistic models of the same data to be combined together by multiplying the probabilities together and renormalizing. An advantage of this method is that each model ('expert') can focus on different aspects of the underlying problem, and that regions of space with high probability mass must satisfy each of the experts, owing to the multiplication of their probabilities.

A product of experts has been successfully used for secondary structure prediction (Chu *et al.*, 2006), where there were separate experts for segmental dependency and strand and helical capping signals. In the present case, we have a distribution for inference of $\mathcal{I}$ given strand pattern $\mathcal{P}_1$, and a distribution for inference of $\mathcal{I}$ given $\mathcal{D}$, a distribution proportional to $\exp[\omega(\mathcal{D},\mathcal{I})]$, described below. Adapting the idea of a product of experts distribution, we use a product of distributions $\mathcal{P}_2(\mathcal{I}|\mathbf{R}) \propto \mathcal{P}_1(\mathcal{I}|\mathbf{R}) \exp[\omega(\mathcal{D},\mathcal{I})]$. When $\mathcal{P}_2(\mathcal{I})$ is large, $\mathcal{I}$ must satisfy both the strand pattern model of $\mathcal{P}_1$ and the CMM $\exp[\omega(\mathcal{D},\mathcal{I})]$. (Formally, $\mathcal{P}_2(\mathcal{I}|\mathbf{R}) = \mathcal{P}(\mathcal{I}|\mathbf{R},\mathcal{D}) \propto \mathbb{P}(\mathcal{I}|\mathbf{R})\mathbb{P}(\mathcal{I}|\mathcal{D})$ and $\mathbb{P}(\mathcal{I}|\mathcal{D}) = \exp[\omega(\mathcal{D},\mathcal{I})]/\sum_i \exp\ [\omega(\mathcal{D},\mathcal{I}_i)]$ where the sum is over the (finite) set of possible $\mathcal{I}_i$.)

Correlated mutation measure, $\exp[\omega(\mathcal{D},\mathcal{I})]$*:* As previously described, $\mathcal{D}(i,j)$ is a measure of how strongly residues in columns i and j co-vary, and a large $\mathcal{D}(i,j)$ suggests residues in columns i and j have co-evolved, and may imply a β-contact between R_i and R_j. This information can be incorporated into the inference as a CMM $\exp[\omega(\mathcal{D},\mathcal{I})]$. The better $\mathcal{I}$ and $\mathcal{D}(i,j)$ fit the larger the value of ω. The formal description of ω follows. We define

$$\emptyset_i = (j : R_j \text{ is a residue in a different strand to residue } R_i).$$

As a concrete example, Figure 5a shows a protein with three strands, residues 3–5, 12–14 and 23–25, where, for example, $\emptyset_4 = \{12, 13, 14, 23, 24, 25\}$ and $\emptyset_{23} = \{3, 4, 5, 12, 13, 14\}$.

In β-sheets, the side chains of residues j and $j \pm 2$ are near each other in space, and so if $\mathcal{D}(i, j \pm 2)$ are large, this may also imply a contact between R_i and R_j. For a particular set of contacts $\mathcal{I}$ and residue R_i, we define the score $\chi(i,\mathcal{I})$ as the mean of the set $\{\mathcal{D}(i,j) : j \in \mathcal{I}_i\}$ where

$$\mathcal{I}_i = \begin{cases} \emptyset_i & \text{if} \nexists j : (i,j) \in \mathcal{I} \\ \bigcup_{j:(i,j)\in\mathcal{I}} \{j-2, j, j+2\} & \text{otherwise} \end{cases}$$

As a concrete example, Figure 5b shows a specific instance of $\mathcal{I}$, and in this case $\mathcal{I}_5 = \emptyset_5, \mathcal{I}_4 = \{12, 14, 16\}$ and $\mathcal{I}_{13} = \{1, 3, 5, 22, 24, 26\}$. The larger $\chi(i,\mathcal{I})$ the better $\mathcal{D}$ and $\mathcal{I}$ fit for residue R_i. However, for different residues R_i, the mean and variance of the set of values $\{\mathcal{D}(i,j) : j \in \emptyset_i\}$ differ wildly and so $\chi(i,\mathcal{I})$ needs to be standardized before being used. For this standardization we take the sample mean μ_i and standard deviation σ_i of the set $\{\mathcal{D}(i,j) : j \in \emptyset_i\}$. So the standardized score, for residue R_i and interaction set $\mathcal{I}$ is then defined as

$$Z(i,\mathcal{I}) = \frac{\chi(i,\mathcal{I}) - \mu_i}{\sigma_i}$$

Defining $\mathcal{I}_{\text{native}}$ as the crystal structure β-contacts defined by DSSP, Figure 6 shows the empirical distribution of $Z(i,\mathcal{I}_{\text{native}})$ over all residues involved in at least one β-contact from the dataset. A much larger mass has positive score than a negative score, implying native contacts have, on average, a larger value for Z.

We then define

$$\omega(\mathcal{D},\mathcal{I}) = \log M \sum_i Z(i,\mathcal{I})$$

where the sum is over all i for which R_i are strands and M = number of sequences in MSA; so that proteins with larger MSA attach more importance to ω.

3 RESULTS AND DISCUSSION

We performed 10-fold cross validation using the same folds as Cheng and Baldi (2005). To estimate posterior probabilities, we used importance sampling. We generated 1 million independent

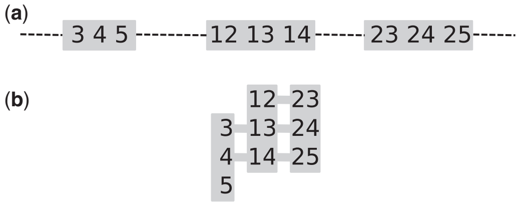


Fig. 5. (**a**) A protein with three strands, residues 3–5, 12–14 and 23–25. (**b**) A specific set of contacts $\mathcal{I} = \{(3,13),(4,14),(12,23),(13,24),(14,25)\}$. See the text for how the standardized score is calculated for this example

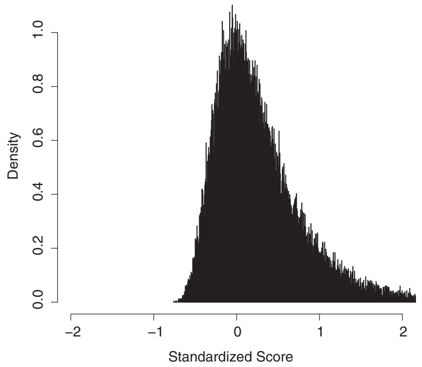


Fig. 6. The empirical distribution of the standardized score, $Z(i, \mathcal{I}_{\text{native}})$ (described in the main text), of all residues involved in at least one β-contact from the dataset

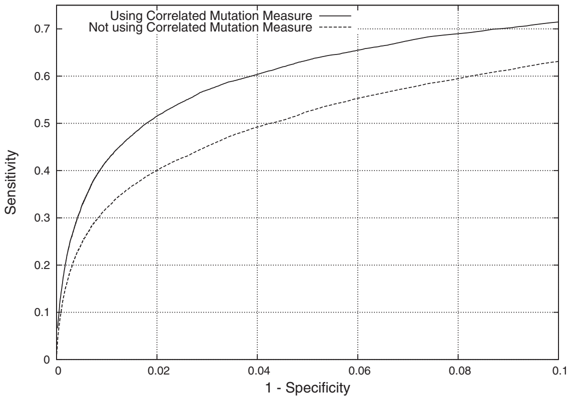


Fig. 7. ROC curves for the posterior both unweighted (dashed) and weighted (solid)

samples from the prior $\mathbb{P}(\mathcal{I}), \{\mathcal{I}\}$ and use these to generate a probability contact map:

$$\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R}) \approx \sum_{\{\mathcal{I}\}} \mathbb{I}[(i,j) \in \mathcal{I}] \frac{\mathbb{P}(\mathbf{R}|\mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathbb{P}(\mathbf{R}|\mathcal{I})}$$

where $\mathbb{I}$ is the indicator function and $\mathbb{P}(\mathbf{R}|\mathcal{I})$ is the likelihood described above (in the case with the CMM we replace $\mathbb{P}(\mathbf{R}|\mathcal{I})$ by $\mathbb{P}(\mathbf{R}|\mathcal{I})\exp[\omega(\mathcal{D},\mathcal{I})]$. See Supplementary Data for further details. We repeated this 50 times and took the mean of the 50 values to generate a single result.

We first quantify the effect of incorporating the CMM into our model. We can take the output of our model and discretize the results, taking as our β-contacts, all (i, j) such that $\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R})$ is larger than a threshold value. Taking different threshold values, Figure 7 shows the receiver operating characteristic (ROC) curve for β-contacts using both the posterior without the CMM $\mathcal{P}_1$ (dashed) and the model using the CMM $\mathcal{P}_2$ (solid). Using the CMM has significantly improved the results. For example, there is a 10% improvement in the number of true positives at the 5% false-positive rate. Figure 7 clearly shows that we have successfully used the evolutionary information, shown to exist in Figures 2 and 6, to improve the prediction of β-contacts.

We can also compare our model with existing β-contact prediction methods. For example, Table 1 shows a comparison with the Neural Network output of the first stage of BetaPro. The results quoted are AUC (Area Under Curve), the true-positive (TP) rate at 5% false positives (FP), TP at the break even point (BEP—when the total number of predicted β-contacts is equal to the true number of β contacts) and the correlation coefficient $\gamma = (\text{TPxTN} - \text{FPxFN})/\sqrt{(\text{TP+FN})(\text{TP+FP})(\text{TN+FN})(\text{TN+FP})}$ at the BEP. This table shows that without the CMM, we produce poorer results than BetaPro. This is to be expected as $\mathcal{P}_1$ is a single sequence method, in contrast to BetaPro that inputs the whole MSA into its neural network. The addition of our CMM improves our method, producing better results than BetaPro.

Unlike some existing models, including BetaPro, our model is completely probabilistic, which enables us to predict both residue-level contacts and strand interactions simultaneously, rather than the latter needing a post processing step. Given strands E_r and E_s, they are defined to be interacting if there exist any β-contact between a residue in strand E_r and a residue in strand E_s. Using our model, we find the following:

$$\mathbb{P}(E_r, E_s \text{ interact}|\mathbf{R}) \approx \sum_{\{\mathcal{I}\}} \mathbb{I}(E_r, E_s \text{ interact}) \frac{\mathbb{P}(\mathbf{R}|\mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathbb{P}(\mathbf{R}|\mathcal{I})}$$

Figure 8 shows the results for two proteins, the N-terminal domain of the yeast HSP90 chaperone [1A4H (left)] and the tetramerization domain of the Shal voltage-gated potassium channel [1NN7 (right)]. For these proteins, our model correctly predicted all strand level interactions and it is interesting to note that for 1NN7, two separate β-sheets are correctly predicted (strands {5,6}, {3,4,1,2} are distinct β-sheets), despite our model not explicitly modelling multiple sheets.

By thresholding the strand interaction probabilities at different values, we can generate a Precision [P = TP/(TP + FP)] versus Recall [R = TP/(TP + FN)] graph for strand interactions, shown in Figure 9. This figure again shows the improvement of the results when we use our CMM.

Table 2 shows a comparison of the strand interactions results for our model, the final output of BetaPro and a MLN (Lippi and Frasconi, 2009). For the comparison, we have only included independent methods and not those such as MLN-2S (Lippi and Frasconi, 2009) or those found in Aydin *et al.* (2011), which are hybrid approaches that combine results from more than one method. The results quoted for our model use the specific probability threshold of 0.45; however, taking the threshold at any

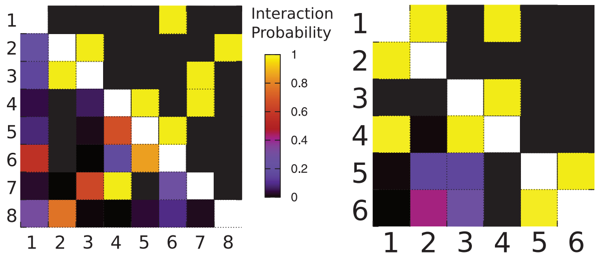


Fig. 8. Contact maps for the strand level for proteins 1A4H (left) and 1NN7 (right). Above the main diagonal, the native (true) strand interactions are shown in yellow, and below the diagonal, $\mathbb{P}(E_r, E_s \text{ interact}|\mathbf{R})$ using $\mathcal{P}_2$ is shown. For protein 1NN7, it is interesting to note that two separate β-sheets are correctly predicted, despite our model not explicitly modelling multiple sheets

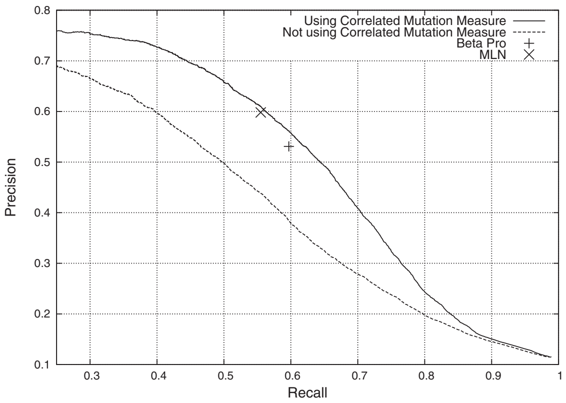


Fig. 9. Precision versus Recall graph for strand interactions. As a comparison, a naive algorithm always pairing adjacent strands yields $P = 0.42$ and $R = 0.50$. The results from the final output of BetaPro and a Markov Logic method (Lippi and Frasconi, 2009) are also displayed for comparison

Table 1. Comparison of the model with the output of BetaPro's Neural Network

Method	AUC	TP at 5% FPR	TP at BEP	γ at BEP
$\mathcal{P}_1$ (no CMM)	0.85	53%	36%	0.34
$\mathcal{P}_2$ (with CMM)	0.89	63%	44%	0.43
BetaPro	0.86	58%	41%	0.4

See text for further details. FPR = false positive rate.

value between 0.25 and 0.63 produces an $F_1 = 2PR/(P+R)$ statistic equal to or above the value found by BetaPro.

The results of our model are clearly better than BetaPro and competitive with MLN. This is an impressive result, as unlike these methods we do not require the additional information of the solvent accessibility of the residues as an input. We also do not require the secondary structure of the non-strand residues, which is important to the MLN method. The only information we use is the maximum entropy-based CMM $\mathcal{D}$ ($\mathcal{P}_1$ is a single sequence method). $\mathcal{D}$ is as useful as providing the entire MSA as a set of 20-dimensional vectors of probabilities as input to a neural or MLN. This may be because providing the columns of the MSA as independent input vectors captures the wrong information; although certain residue pairs are more likely to form β-contacts (for example, pairs of hydrophobic residues in the core of a protein), the individual pairing preferences are not especially strong, and proteins do not seem to have strong evolutionary pressure to maintain favourable pairings between strands (Mandel-Gutfreund *et al.*, 2001).

Table 2. Comparison of strand level statistics of our model ($\mathcal{P}_2$), the final output of BetaPro and the MLN method of Lippi and Frasconi (2009)

Statistic	$\mathcal{P}_2$	BetaPro	MLN
P	61.0	53.1	59.8
R	55.4	59.7	55.5
F_1	58.1	56.2	57.6
γ	0.532	0.508	0.528
Chains with $F_1 \geq 70.0$	35.0	31.7	33.7

Apart from γ, statistics are shown as percentages.

Also, just considering the specific residue types, rather than how they co-vary, suffers from the problem of transitivity: if E_r is paired with both E_s and E_t, then it is often the case E_t and E_s themselves contain residues with favourable pairings, as they both favourably interact with E_r.

For our method to be useful for proteins with unknown structure, it is important to test our method with predicted secondary structure. In the Supplementary Data, we have presented results for the CASP 2010 set of proteins using both known and predicted strand structure, and in both cases our method compares favourably with BetaPro.

4 CONCLUSION AND FURTHER WORK

In this article, we have used a statistical machine learning approach known as contrastive divergence to efficiently calculate a Maximum Entropy distribution that models the evolutionarily related family of a protein and have used this to calculate a CMM to predict residue contacts. We have coupled this measure to a probabilistic model of β-strand interactions to produce a state-of-the-art β-contact predictor that can be used even if a poor quality or no MSA is available. The current focus of our work is to allow joint inference of β-contacts and secondary structure by incorporating a semi-segmental Markov model to model the secondary structure of proteins (Chu *et al.*, 2006; Schmidler *et al.*, 2000).

Unlike other recent CM studies, we have focused on proteins that do not necessarily have large enough MSAs to enable full tertiary structure determination using a CM approach. However, our strand interaction prediction can be incorporated into a tertiary structure prediction method. For example, our previously published work describes a coarse-grained protein model that uses a physically meaningful energy function, biased by a harmonic potential on β-contacts to enable the protein to fold (Burkoff *et al.*, 2012; Podtelezhnikov and Wild, 2008). Using our strand prediction method to predict β-contacts enables this model to be used for protein tertiary structure prediction.

Further details and specific examples are shown in the Supplementary Data.

ACKNOWLEDGEMENT

We thank John Skilling for helpful discussions.

Funding: We acknowledge support from the Leverhulme Trust (grant F/00 215/BL to N.S.B, C.V. and D.L.W.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search proteins. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashkenazy,H. *et al.* (2009) Optimal data collection for correlated mutation analysis. *Proteins*, **74**, 545–555.

Aydin,Z. *et al.* (2011) Bayesian models and algorithms for protein β-sheet prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **8**, 395–409.

Balakrishnan,S. *et al.* (2011) Learning generative models for protein fold families. *Proteins*, **79**, 1061–1078.

Burger,L. and van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633–51.

Burkoff,N.S. *et al.* (2012) Exploring the energy landscapes of protein folding simulations with bayesian computation. *Biophysical. J.*, **102**, 878–886.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21**, i75–i84.

Chu,W. *et al.* (2006) Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **3**, 98–113.

Dekker,J.P. *et al.* (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.

Duarte,J.M. *et al.* (2010) Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, **11**, 283.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, **32**, 1792–1797.

Gobel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Hinton,G.E. (1999) Products of experts. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks*, Vol. 1. University of Edinburgh, UK, pp. 1–6.

Hinton,G.E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Comput.*, **14**, 1771–1800.

Hopf,T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.

Hubbard,T.J. and Park,J. (1994) Use of β-strand interaction pseudo potentials in protein structure and modelling. In: *Proceedings of the 27th Hawaii Int'l Conf. System Sciences*. Maui, HI, USA, pp. 336–344.

Jaynes,E.T. (2007) *Probability Theory: The Logic of Science*. CUP, Cambridge, UK.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kortemme,T. *et al.* (1998) Design of a 20-amino acid, three-stranded β-sheet protein. *Science*, **281**, 253–256.

Lee,B.-C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, **17**, 282–283.

Lippi,M. and Frasconi,P. (2009) Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. *Bioinformatics*, **25**, 2326–2333.

Mandel-Gutfreund,Y. *et al.* (2001) Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J. Mol. Biol.*, **305**, 1145–1149.

Marks,D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Merkel,J.S. and Regan,L. (2000) Modulating protein folding rates in vivo and in vitro by side chain interactions between the parallel beta strands of green fluorescent protein. *J. Biol. Chem.*, **275**, 29200–29206.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA*, **108**, E1293–E1301.

Olmea,O. and Valencia,A. (1997) Improving contact prediction by the combination of correlated mutations and other sources of sequence information. *Fold. Des.*, **2**, S25–S31.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

Podtelezhnikov,A.A. and Wild,D.L. (2008) Crankite: a fast polypeptide backbone conformation sampler. *Source Code Biol. Med.*, **3**, 12.

Podtelezhnikov,A.A. and Wild,D.L. (2009) Reconstruction and stability of secondary structure elements in the context of protein structure prediction. *Biophys. J.*, **96**, 4399–4408.

Rajgaria,R. *et al.* (2010) Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. *Proteins*, **78**, 1825–1846.

Ruczinski,I. *et al.* (2002) Distribution of beta sheets in proteins with application to structure prediction. *Proteins*, **48**, 85–97.

Saqi,M.A.S. *et al.* (1999) Protein analyst—a distributed object environment for protein sequence and structure analysis. *Bioinformatics*, **15**, 521–522.

Sathyapriya,R. *et al.* (2009) Defining an essence of structure determining residue contacts in proteins. *PLoS Comput. Biol.*, **5**, e1000584.

Schmidler,S.C. (2002) Statistical models and monte carlo methods for protein structure prediction. In: PhD Thesis, Stanford University, Stanford, CA, USA.

Schmidler,S.C. *et al.* (2000) Bayesian segmentation of protein secondary structure. *J. Comput. Biol.*, **7**, 232–248.

Smith,C.K. and Regan,L. (1995) Guidelines for protein design: the energetics of β sheet side chain interactions. *Science*, **270**, 980–982.

Sonnhammer,E.L.L. *et al.* (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, **28**, 405–420.

Steward,R.E. and Thornton,J.M. (2002) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins Struct. Funct. Genet.*, **48**, 178–191.

Sułkowska,J. *et al.* (2012) Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA*, **109**, 10340–10345.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold. Des.*, **2**, 295–306.

Weigt,M. *et al.* (2008) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.

1 Supplementary Data

1.1 Calculating the Correlated Mutation Measure

1.1.1 Maximum Entropy Distribution

Given a MSA containing M sequences for a protein of length N, we define $f_i(A_i)$ as the observed frequency of residue A_i occuring in position i of the MSA and $f_{ij}(A_i, A_j)$ as the observed frequency of both residue A_i occuring in position i and residue A_j occuring in position j of the MSA. Given any sequence $\mathbf{A} = A_1, A_2, \ldots, A_N$, we model the probability of it occurring in the MSA by a distribution $P(\mathbf{A}) = P(A_1, A_2, \ldots, A_N)$. However, there are q^N different sequences (where $q = 21$ is the size of the alphabet of amino acids – in these models the 'gap' residue is treated as a 21st amino acid type) and only $M \ll q^N$ sequences in the MSA. The sparsity of the data and the number of sequences imply it is impractical for detailed use. However, we would like our model to match the emiprical low-order moments given by the MSA. Specifically we would like

$$P_i(A_i) = f_i(A_i) \quad \text{and} \quad P_{ij}(A_i, A_j) = f_{ij}(A_i, A_j) \tag{1}$$

where $P_i(.)$ is the marginal distribution for position i and $P_{ij}(.,.)$ is the (joint) marginal distribution of positions i and j. In this work we have not added pseudo-counts or weighted sequences.

Among the valid distributions P satisfying these constraints, using the maximum entropy principle [9] we favour, $\mathfrak{P}$, the distibution which has maximum entropy, S:

$$\mathfrak{P} = \text{argmax}_P\{S(P)\}\equiv\text{argmax}_P\{-\sum_{\mathbf{A}} P(\mathbf{A}) \log(P(\mathbf{A}))\}$$

and solving this optimization problem using Lagrange multipliers leads to the distribution

$$\mathfrak{P}(A_1, \ldots, A_N) \propto \exp\left(\sum_{1\leq i \leq j \leq N} -e_{ij}(A_i, A_j) + \sum_{1 \leq i \leq N} h_i(A_i)\right) \tag{2}$$

for some pair-interaction energies $e_{ij}(A_i, A_j)$ and local fields $h_i(A_i)$ [17].

The number of parameters specificed by equation 2 is $N(N-1)q^2/2 + Nq$ yet there are only $N(N-1)(q-1)^2/2 + N(q-1)$ free parameters because the constraints from (1) are not independent, for example $\sum_{A_j} f_{ij}(A_i, A_j) = f_i(A_i)$. To uniquely specify a solution requires the imposition of so-called gauge conditions. We follow the approach of [17](Supporting Information); since we are trying to estimate the coupling between residue pairs it intuitively makes sense to capture as much as possible by the local fields h_i and to leave only the *essential* pairwise contributions in the interaction energies e_{ij}. Formally this can be achieved by imposing the gauge constraints

$$\sum_{A_j} e_{ij}(A_i, A_j) = \sum_{A_i} e_{ij}(A_i, A_j) = 0$$

for all i, j, A_i, A_j. Finally to fully determine the system we fix

$$\sum_{A_i} h_i(A_i) = C_i$$

for all i, where C_i is a constant, chosen for simplicity as $\sum_{A_i} \ln(f_i(A_i))$.

To drastically reduce the dimensionality of the problem, if $f_i(A_i) = 0$ for some i and A_i, then we do not need to use $h_i(A_i)$ or $e_{ij}(A_i, A_j)$ for all j and A_j (in effect we are setting $h_i(A_i) = -\infty$). This reduces both the number of independent variables and the number of constraints.

1.1.2 Contrastive Divergence

Due to the large number of e_{ij}, it is a significant computational challenge to compute $\mathfrak{P}$ and many different strategies have previously been employed. These include Monte Carlo samping [15], message passing [17], perturbation methods [4] and mean-field approximations [13, 12]. In this work we take a different approach; we use contrastive divergence, a statistical machine learning technique introduced by Hinton [7]. This work represents the first application of contrastive divergence to the modelling of protein multiple sequence alignments. We have chosen contrastive divergence because Equation 2 defines a (restricted) Boltzmann machine [8, 16] and contrastive divergence is a technique which has been developed to efficiently learn the parameters of Boltzmann machines.

Given a MSA, we replicate it to ensure it contains $M \geq 5000$ sequences. We denote this ensemble by $\mathcal{E}_\infty$ and view it as a single data point. For a given set of $\theta^m = \{e_{ij}^{(m)}, h_i^{(m)}\}$ we define the distribution π_m as

$$\pi_m(\mathbf{A}) = \pi_m(A_1, \dots, A_N) \propto \exp\left(- \sum_{1 \leq i \leq j \leq N} e_{ij}^{(m)}(A_i, A_j) + \sum_{1 \leq i \leq N} h_i^{(m)}(A_i) \right)$$

and we can define $\pi_m(\mathcal{E}_\infty) = \prod_{X \in \mathcal{E}_\infty} \pi_m(X)$. Using Maximum Likelihood (ML) optimization, with learning rate η, we would update θ using

$$\theta^{m+1} = \theta^m + \eta \frac{\partial}{\partial \theta} \ln \pi_m(\mathcal{E}_\infty) = \theta^m + \eta \left[\frac{\partial \ln \pi_m(\mathcal{E}_\infty)}{\partial \theta} - \left\langle \frac{\partial \ln \pi_m(\mathcal{E})}{\partial \theta} \right\rangle \right]$$

where $\left\langle \frac{\partial \ln \pi_m(\mathcal{E})}{\partial \theta} \right\rangle$ is the average with respect to all possible data points $\mathcal{E}$. This *ensemble* average is extremely computationally expensive to compute; typically, a Markov Chain (MC) would have to be run to equilibrium for each ML iteration.

Contrastive divergence is an alternative method whereby one runs a short MC chain, starting from $\mathcal{E}_\infty$ and using the stationary distribution π_m, ending at state $\mathcal{E}_m$. The approximation

$$\left\langle \frac{\partial \ln \pi_m(\mathcal{E})}{\partial \theta} \right\rangle \approx \frac{\partial \ln \pi_m(\mathcal{E}_m)}{\partial \theta}$$

is then used. For a full discussion of the theory behind contrastive divergence see [7]. Using a simple gradient descent update scheme does not preserve the gauge conditions, so we follow the update scheme described in [17](Supporting Information). The complete algorithm is described by Algorithm 1 and the independent MC sampler moves are described by Algorithm 2.

1.1.3 Correlated Mutation Measure

Once we have the values of e_{ij} and h_i, we need a measure which quantifies the coupling between residues R_i and R_j. In this work we use $\mathcal{D}(i, j)$ which is defined as

$$\mathcal{D}(i, j) = \sum_{A_i, A_j} P_{ij}^D(A_i, A_j) \log \frac{P_{ij}^D(A_i, A_j)}{f_i(A_i) f_j(A_j)}$$

where

$$P_{ij}^D(A_i, A_j) \propto f_i(A_i) f_j(A_j) \exp(-e_{ij}(A_i, A_j)).$$

For $\{i, j\}$ this is the relative entropy between the independent site joint distribution $f_i(A_i) f_j(A_j)$ and the independent site joint distribution moderated by *only* this pair's interaction energies. The larger $\mathcal{D}(i, j)$, the greater the coupling between i and j and this may indicate co-evolution and a residue contact.

Many measures are possible, for example, $\sum_{A_i, A_j} [e_{ij}(A_i, A_j)]^2$, and these measures are found to contain essentially the same information [17]. In previous work the *direct information* [12, 13, 17] was used, which has

Algorithm 1 Procedure to calculate $\mathfrak{P}$ using contrastive divergence (with an independent MC sampler)

```
$\mathcal{E}_\infty \leftarrow$ MSA (copied to ensure 5000 sequences in ensemble).
$e_{ij}^{(0)}(A_i, A_j) \leftarrow 0$, $h_i^{(0)}(A_i) \leftarrow f_i(A_i)$
$m \leftarrow 0$, $q \leftarrow 21$
loop
    $\pi_m \leftarrow$ Maximum Entropy Distribution with $e_{ij}^{(m)}$ and $h_i^{(m)}$
    $\mathcal{E}_m \leftarrow \mathcal{E}_\infty$
    for $n = 1 \rightarrow 5000$ do
        Propose MC move $\mathcal{E}_m \rightarrow \mathcal{E}_{m'}$ (see Algorithm 2)
        $U \sim \mathcal{U}[0, 1]$
        if $U \leq \min\{1, \frac{\pi_m(\mathcal{E}_{m'})}{\pi_m(\mathcal{E}_m)}\}$ then
            $\mathcal{E}_m \leftarrow \mathcal{E}_{m'}$
        end if
    end for
    Update $e_{ij}$ and $h_i$:
        $P_i(A_i), P_{ij}(A_i, A_j) \leftarrow$ empirical first and second order moments of $\mathcal{E}_m$
        $\Delta h_i(A_i) \leftarrow f_i(A_i) - P_i(A_i)$
        $\Delta e_{ij}(A_i, A_j) \leftarrow [f_{ij}(A_i, A_j) - P_{ij}(A_i, A_j)] - \frac{f_i(A_i)+f_j(A_j)-P_i(A_i)-P_j(A_j)}{q}$
        $h_i^{(m+1)}(A_i) \leftarrow h_i^{(m)}(A_i) + \eta\Delta h_i(A_i)$
        $e_{ij}^{(m+1)}(A_i, A_j) \leftarrow e_{ij}^{(m)}(A_i, A_j) + \eta\Delta e_{ij}(A_i, A_j)$
        $m \leftarrow m + 1$
end loop
```

Algorithm 2 Proposing MC move $\mathcal{E}_m \rightarrow \mathcal{E}_{m'}$

```
Uniformly choose a sequence from $\mathcal{E}_m$: $S$
Uniformly choose a starting column: $s$
Uniformly choose block size $k$ from the set $\{0, 1, 2, 3, 4\}$,
for $i = 0 \rightarrow k$ do
    Change residue $s + i$ of sequence $S$, choosing uniformly over alphabet of amino acids which are found in column $s + i$ of $\mathcal{E}_\infty$.
end for
The MC chain is now in state $\mathcal{E}_{m'}$
```

Notes:
The standard independent Metropolis-Hasting acceptance criteria (described in Algorithm 1) is then used. If $s + i$ is greater than L, the number of residues in each sequence then we wrap around back to column 0.

the advantage of being gauge invariant. From a theoretical viewpoint, however, since our gauge constraint has been chosen so that the h_i explain as much as possible, what is left in e_{ij} is, in some sense, the essential pairwise interactions. Numerically, the direct information was also used for this work but produced slightly poorer prediction results. This is shown in Figure 1 where, for each protein, the contact ratio (as defined in the caption for Figure 2 (Top) in the main text) is plotted using both the the new correlation measure $\mathcal{D}$ and the original direct information. As expected both measures contain similar information (the correlation coefficient is 0.948) yet the direct information is slightly poorer: the linear regression line shows (Direct Information) = $0.9(\mathcal{D}) + 0.0034$. Finally, unlike the direct information, the new correlation measure has the computational advantage of not having to calculate $\mathcal{O}(N^2)$ additional fields after the maximum entropy distribution has been calculated.

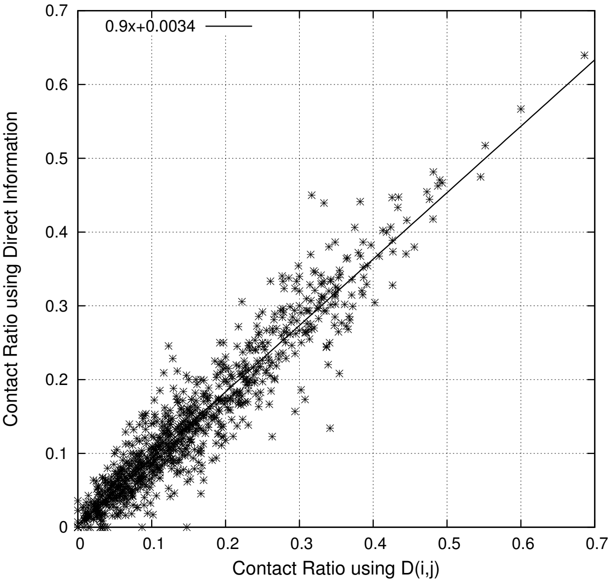


Figure 1: The contact ratio using both $\mathcal{D}$ (x-axis) and the direct information (y-axis) is plotted for each of the 916 proteins in the dataset. Both correlation measures contain essentially the same information (the correlation coefficient is 0.948) yet the new measure performs slightly better, as shown by the linear regression line.

Noise associated with correlated mutation analysis: There are distinct types of noise associated with contact prediction from correlated mutation analysis. The first is independent of the size of the MSA and concerns *direct* and *indirect* contacts as described in Section 2.2 of the main text. As a specific example, if residues X and Y are in contact and correlated and residues Y and Z are in contact and correlated, X and Z, which are not in contact, are likely to be correlated. A pairwise correlation measure would be unable to distinguish between the actual contacts and the indirect correlations.

The maximum entropy formalism is a global (rather than pairwise) method which aims to find the minimally constrained distribution (as defined by entropy) which satisfies all observed correlations. For example, in the above case, a large interaction energy between X and Y and between Y and Z would be enough to explain the correlation between X and Z.

There are other causes of noise, for example, in the case of a homodimer made up of two identical monomers A and A'. Chain A residue X could be in contact with and hence co-evolve and be correlated with chain A', residue Y. Correlated mutation analysis could then seem to imply that residues X and Y are in contact within the monomer, which is not the case. This is an issue for all correlated mutation measures, although the maximum entropy formalism performs strongly despite this noise. In this paper, for the case of β-contacts, we have tried to mitigate this noise, as well as the noise occurring due to lack of data, by coupling the correlation measure to a strongly structured strand model which only allows plausible β-contacts. As a simple example, if residues (i, j) and $(i+1, j+1)$ are β-contacts, residues $(i+1, j-1)$ cannot be a β-contact even if positions $i+1$ and $j-1$ of the MSA are highly correlated.

1.2 Empirical Bayes Parameter Optimization

We weight the likelihood with a gamma distribution on the number of β-contacts $(|\mathcal{I}|^{(u|E|-1)} \exp(-v|\mathcal{I}|))$ for the reasons outlined in the main text. To optimize the parameters u and v we used an empirical Bayes technique. We set the mean and variance of the gamma distribution as $\mu|E|$ and $\sigma^2|E|$ where $|E|$ is the number of β-residues. For a fixed μ and σ^2 we have

$$v = \frac{\mu}{\sigma^2} \quad \text{and} \quad u = \frac{\mu^2}{\sigma^2}$$

Fixing $\{\mu, \sigma^2\}$ we generated 3 independent sets of results using the same 10-fold cross validation procedure described in the main text (each set of results was the mean of 20 sets of 250000 samples per protein). For each set of results we calculated the maximum value of $F_1 = 2PR/(P+R)$, the harmonic mean of Precision and Recall for strand interactions, and took the mean of these three maximum values. We then repeated this for 120 other pairs of $\{\mu, \sigma^2\}$. Figure 2 shows the results. The results seem to show that a large range of $\{\mu, \sigma^2\}$ produce statistically equivalent results and we chose $\mu = 1.5$ and $\sigma^2 = 0.125$, which implies $u = 18$ and $v = 12$. Although we chose the maximum $F_1 = 2PR/(P+R)$ value as our statistic to compare, the same behaviour was found when using other statistics such as the correlation coefficient γ.

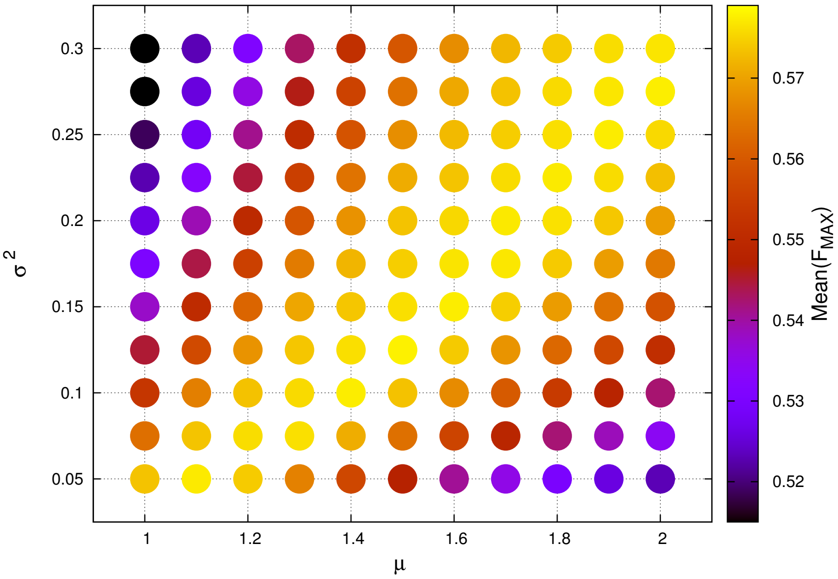


Figure 2: The maximum value of $F_1 = 2PR/(P+R)$ for different values of $\{\mu, \sigma^2\}$. See the text for more details.

1.3 Importance Sampling:

Following [3] we use importance sampling to estimate properties of the posterior distribution. We draw 1 million samples $\{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ from our proposal distribution, the prior: $\mathbb{P}(\mathcal{I})$, and use the standard importance sampling weights:

$$W_i = \frac{\mathbb{P}(\mathcal{I}_i|\mathbf{R})}{\mathbb{P}(\mathcal{I}_i)} = \frac{\mathbb{P}(\mathbf{R}|\mathcal{I}_i)}{Z}$$

where

$$Z = \sum_{\mathcal{I}} \mathbb{P}(\mathbf{R}|\mathcal{I})\mathbb{P}(\mathcal{I}) \approx \sum_{\{\mathcal{I}_i\}} \mathbb{P}(\mathbf{R}|\mathcal{I}_i) \tag{3}$$

Given a property of the posterior $h(\mathcal{I})$ we can estimate its expectation:

$$\mathbb{E}(h(\mathcal{I})) \approx \sum_{\{\mathcal{I}_i\}} h(\mathcal{I})W_i$$

and we take the mean of 50 independent estimates of $\mathbb{E}(h(\mathcal{I}))$ to produce the results shown in the paper. Since such a large number of samples is taken and the fact that the prior itself is extremely informative (especially for the posterior which does not include the correlated mutation measure, $\mathcal{P}_1$), we find the approximation in Equation 3 is valid.

1.4 CASP 2010 Dataset

We have also benchmarked our method on the CASP 2010 set of proteins. We removed those proteins which have fewer than 10 β-residues and we also removed protein T0543, an 887 residue protein, leaving a total of 92 proteins. We followed the same procedure described in the paper to autogenerate MSAs. Figure 3 shows a comparison between the number of sequences in the MSAs of the CASP dataset and the BetaPro 916 dataset shown in the paper. There are more CASP targets with larger MSAs, for example 30% of the CASP dataset have at least 1000 sequences in their MSA as opposed to 22% of the BetaPro 916 dataset. Nevertheless, similarly to the dataset in the main paper, a third of the CASP dataset still have fewer than 200 sequences in their MSA and so this data set also provides an additional test of our method when only small MSAs are available.

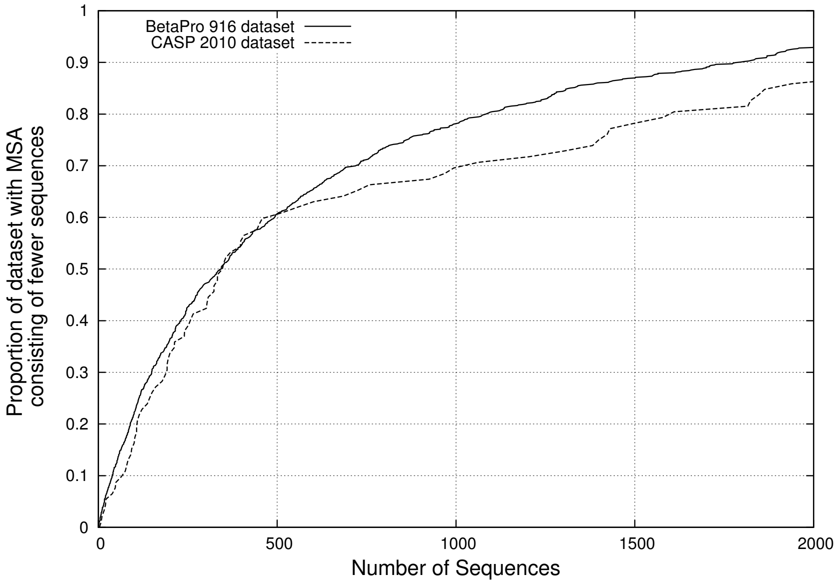


Figure 3: A comparison between the number of sequences in the MSAs of the 916 dataset from the main paper and the CASP 2010 92 proteins.

1.4.1 Known Secondary Structure

Using constrastive divergence we generated our correlated mutation statistic $\mathcal{D}$, and in the first instance we tested our strand predictor program using the secondary structure assignments from DSSP [11]. For BetaPro we also used solvent accessibilty derived from DSSP. The results were similar to those of the BetaPro 916 dataset, for example, Figure 4 and Table 1 show the strand level results of our method compared to BetaPro.

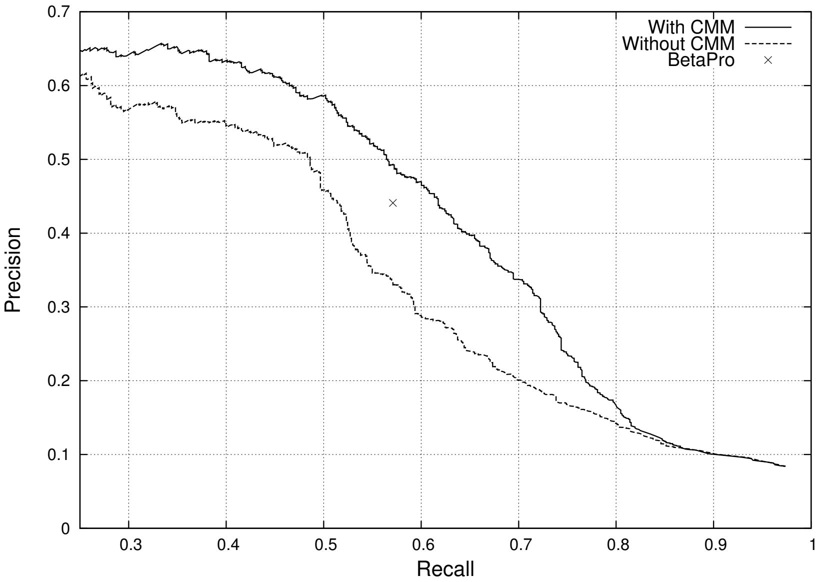


Figure 4: Precision versus Recall graph for strand interactions on the CASP dataset using Correlated Mutation Measure (CMM): $\mathcal{P}_2$, and not using CMM: $\mathcal{P}_1$. Compare to Figure 9 in the main text.

Table 1: Comparison of strand level statistics of our model ($\mathcal{P}_2$) with threshold 0.45 and the final output of BetaPro for the CASP 2010 dataset. Apart from γ, statistics are shown as percentages. Compare with Table 2 in the main text.

	$\mathcal{P}_2$	BetaPro
P	53.1	44.1
R	54.2	57.1
F_1	53.7	50.0
γ	0.495	0.451
Chains with $F_1 \geq 70.0$	32.6	32.6

1.4.2 Predicted Secondary Structure

We have also tested our method with *predicted* secondary structure. We predicted strand residues using both Jpred [5, 6] and PsiPred [1, 10], taking a residue to be part of a strand if at least one of the methods predicted it to be a strand residue. For the CASP dataset, 81.2% of the strand residues were correctly predicted and of those predicted as strands 79.6% were actually strands. For BetaPro we also predicted helices using the same criterion and predicted solvent accessibility using Jpred. Figure 5 shows the ROC curve at the residue level[1] of the CASP dataset using predicted secondary structure. It shows that in the region of few false positives, we perform slightly better than Betapro. It is important to note that this graph cannot be compared to the ROC curve in the main paper (Figure 7) because in this case there are many more true negatives, as all pairs of residues could be in β-contact as opposed to only the known β-residues.

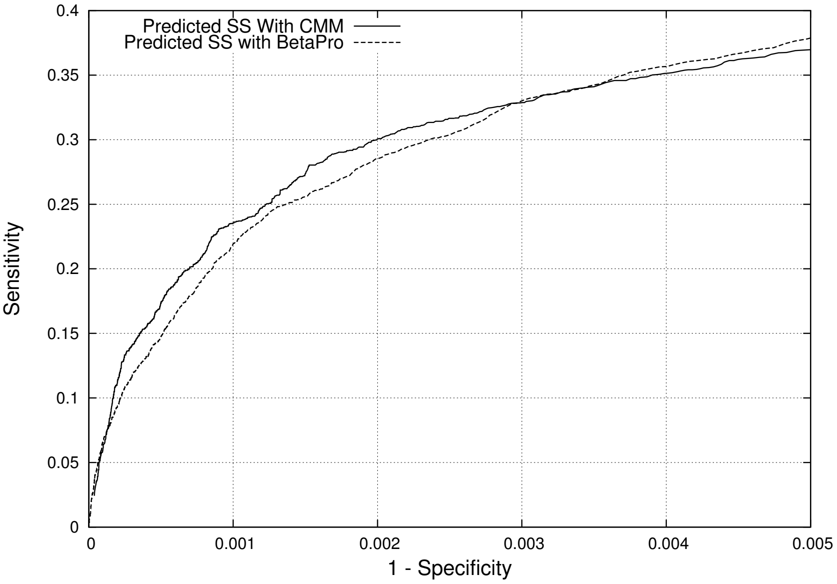


Figure 5: ROC curve for residue level β-contacts using correlated mutation measure, compared to the neural network output of BetaPro. It is important to note that this graph cannot be compared to the ROC curve in the main paper (Figure 7) because in this case there are many more true negatives, as all pairs of residues could be in β-contact as opposed to only the known β-residues.

1.5 Tertiary Structure Prediction

As described in the main text, our β-contact predictions can be incorporated into a tertiary structure prediction method. Our previously published software, CRANKITE, describes a coarse-grained protein model which uses a physically meaningful energy function biased by a harmonic potential on β-contacts to fold a protein [14, 2]. CRANKITE requires protein secondary structure and β-contacts as an input. Using our strand prediction method to predict β-contacts enables our coarse-grained model to be used for protein tertiary structure prediction. A full account of the physical model and statistical sampling method can be found in [14, 2][2].

Figures 6 and 7 are examples of tertiary structure prediction of the 2010 CASP targets T0622 (a 122 residue protein with 968 sequences in its MSA) and T0594 (a 140 residue protein with only 258 sequences in its MSA). These figures include the primary sequence, native (crystal structure) and predicted secondary structure. The predicted secondary structure uses the same consensus method described in Section 1.4.2. The figures also include the native residue level β-contacts and the contacts predicted using our β-strand

[1] It no longer makes sense to describe strand level interactions as there is no clear correspondance between predicted strands and actual strands in the protein.

[2] An additional energy term constraining the radius of gyration of the centre of masses of the distinct helices and strands has been developed since our last publication, this energy term allows secondary structure elements to form, yet ensures the overall conformation is compact.

model (analogous to the strand level heat maps of Figure 8 in the main text) using the predicted secondary structure. Finally the native (crystal) structure of the protein is shown along with example low energy structures found by CRANKITE together with their RMSDs from the crystal structure.

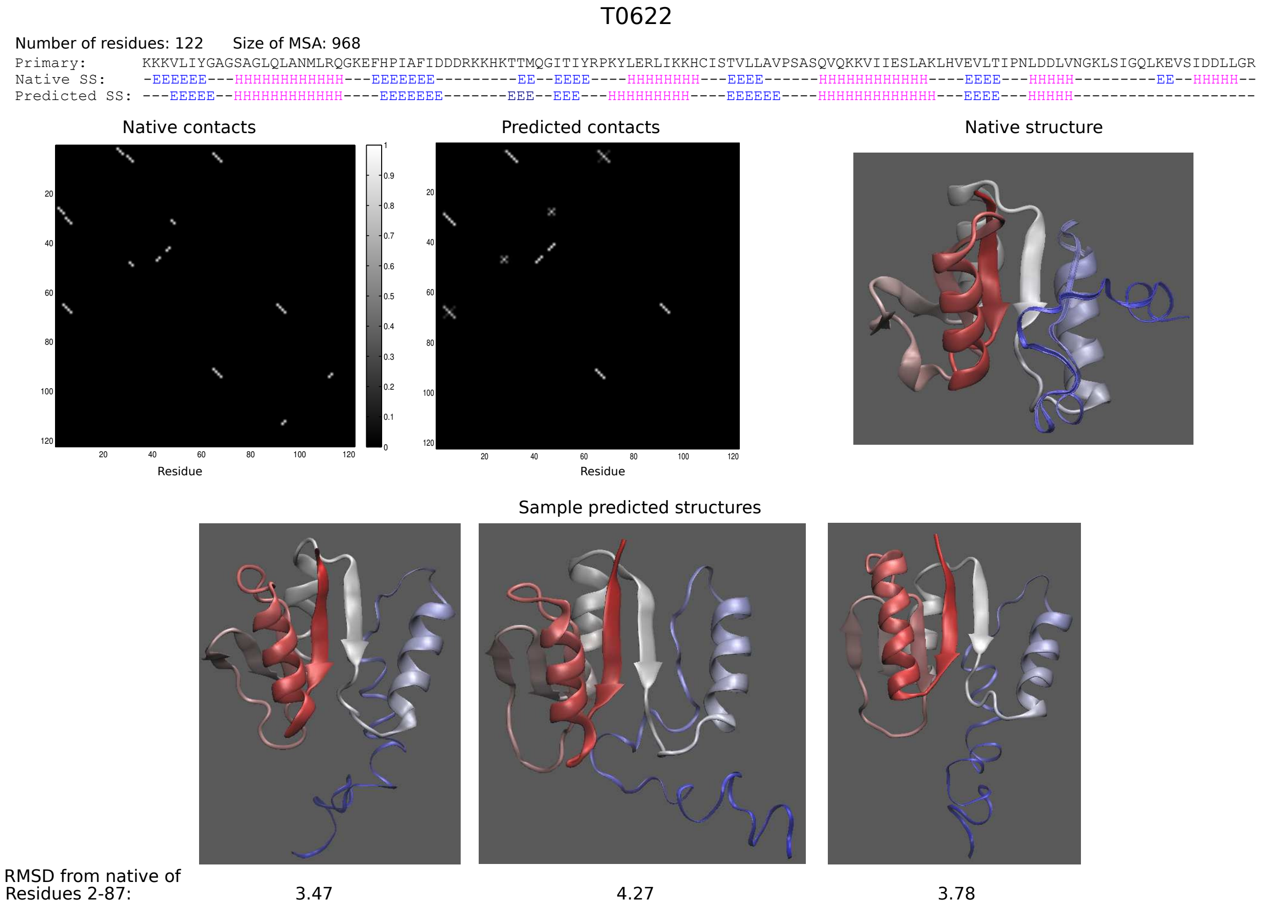

T0622
Number of residues: 122 Size of MSA: 968
Primary: KKKVLIYGAGSAGLQLANMLRQGKEFHPIAFIDDDRKKHKTTMQGITIYRPKYLERLIKKHCISTVLLAVPSASQVQKKVIIESLAKLHVEVLTIPNLDDLVNGKLSIGQLKEVSIDDLLGR
Native SS: -EEEEEE---HHHHHHHHHHHH---EEEEEEE---------EE--EEEE----HHHHHHHH---EEEE------HHHHHHHHHHHH----EEEE---HHHHH---------EE--HHHHH--
Predicted SS: ---EEEEE--HHHHHHHHHHHH----EEEEEEE-------EEE--EEE---HHHHHHHHH----EEEEEE----HHHHHHHHHHHHH---EEEE---HHHHH--------------------
Native contacts
Predicted contacts
Native structure
20
40
60
80
100
120
Residue
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Sample predicted structures
RMSD from native of Residues 2-87:
3.47
4.27
3.78

Figure 6: Tertiary structure prediction of 2010 CASP target T0622, a 122 residue protein with 968 sequences in its MSA. Top: The protein primary sequence, native (crystal structure) secondary structure (ss) and predicted secondary structure. Middle: Actual residue β-contacts, β-contact prediction using our model and the native structure of the protein. Bottom: Example low energy structures found using CRANKITE. The C-terminal tail of the protein, the blue portion, has been rendered using 'ribbon' view so that the topology of the rest of the protein can be clearly seen. Note for this protein the contact between residue 28 and 47 was used as an input for CRANKITE but it is unclear from the prediction whether this strand interaction is parallel or antiparallel. Therefore, only this contact was used for this strand interaction and CRANKITE found the most favourable direction itself.

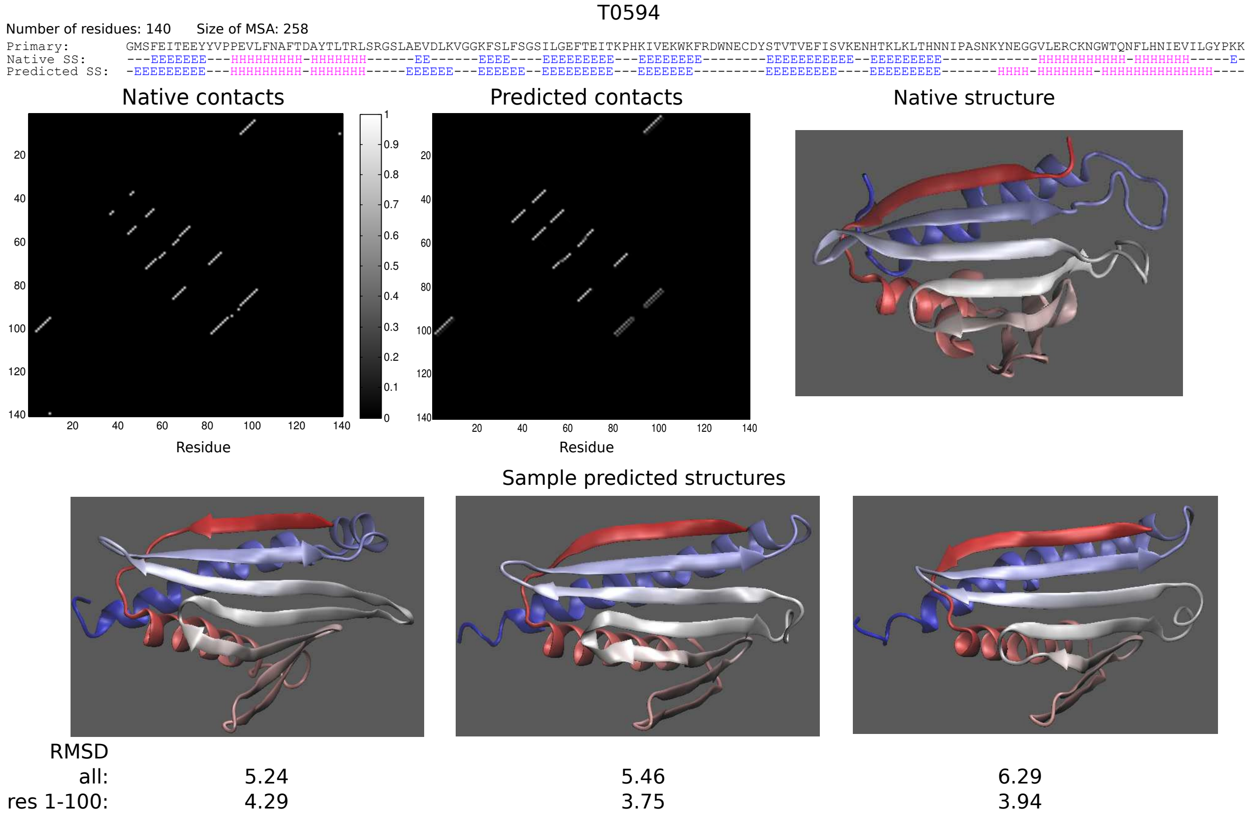


Figure 7: Tertiary structure prediction of 2010 CASP target T0594, a 140 residue protein with only 258 sequences in its MSA. Top: The protein primary sequence, native (crystal structure) secondary structure (ss) and predicted secondary structure. Middle: Actual residue β-contacts, β-contact prediction using our model and the native structure of the protein. Bottom: Example low energy structures found using CRANKITE. The β-contact prediction produced two possible residue level interactions for the final two strands, the one with larger probability was chosen as input for CRANKITE.

References

[1] D. W. Buchan, S. M. Ward, A. E. Lobley, T. C. Nugent, K. Bryson, and D. T. Jones. Protein annotation and modelling servers at University College London. *Nucl. Acids Res.*, 38 Suppl:W563–8, 2010.

[2] N. S. Burkoff, C. Várnai, S. A. Wells, and D. L. Wild. Exploring the Energy Landscapes of Protein Folding Simulations with Bayesian Computation. *Biophysical J.*, 102(4):878–886, 2012.

[3] W. Chu, Z. Ghahramani, A. Podtelezhnikov, and D. L. Wild. Bayesian Segmental Models with Multiple Sequence Alignment Profiles for Protein Secondary Structure and Contact Map Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):98–113, 2006.

[4] S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc. Natl. Acad. Sci*, 106:14058–14062, 2009.

[5] C. Cole, J. D. Barber, and G. J. Barton. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, 36 (suppl. 2):W197–201, 2008.

[6] J. A. Cuff and G. J. Barton. Application of Enhanced Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction. *PROTEINS: Structure, Function and Genetics*, 40:502–51, 2000.

[7] G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14:1771–1800, 2002.

[8] G. E. Hinton and T. J. Sejnowski. *Learning and relearning in Boltzmann machines. In: Parallel Distributed Computing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations.* MIT Press, 1986.

[9] E. T. Jaynes. *Probability Theory: The Logic of Science.* CUP, 2007.

[10] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*, 292:195–202, 1999.

[11] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[12] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One*, 6(12):e28766, 2011.

[13] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. Plus*, 108(49):E1293–E1301, 2011.

[14] A. A. Podtelezhnikov and D. L. Wild. Crankite: A fast polypeptide backbone conformation sampler. *Source Code Biol. Med.*, 3:12, 2008.

[15] E. Schneidman, J. M. Berry 2nd R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in neural population. *Nature*, 440:1007–1012, 2006.

[16] P. Smolensky. *Information processing in dynamical systems: Foundations of harmony theory. In: Parallel Distributed Computing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations.* MIT Press, 1986.

[17] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA*, 106(1):67–72, 2008.

Chapter 5

Nested Sampling with Molecular Dynamics

In Chapter 2 we applied the nested sampling algorithm to a coarse-grained protein model specifically designed to allow the implementation of efficient MC crankshaft moves. This is due to the fixed bond lengths and valence angles (with the exception of the C_α atom). However, these degrees of freedom are not fixed in proteins, and hence many standard all-atom force fields allow bond stretching and angle bending.

For these models MC sampling is not typically used. For example, it is not included with the popular bimolecular modelling package, Amber. Therefore, in order for nested sampling to become widely used within the computational biophysics community, it would be advantageous to implement nested sampling within a MD framework.

Skilling has recently developed Galilean nested sampling (GNS), a variant of nested sampling whereby each degree of freedom is given a 'velocity' and new samples are generated by using these velocities to evolve existing samples along trajectories. Rather than Hamiltonian (i.e. canonical or microcanonical) trajectories, samples follow *Galilean* trajectories, a novel exploration procedure designed to sample phase space uniformly. This ensures GNS retains the advantages of nested sampling at first order phase transitions.

GNS is ideally suited to sample atomistic systems and in this chapter we implement the algorithm as a wrapper to the Amber MD package.[1] We demonstrate the validity of GNS by sampling the small peptide alanine dipeptide, a common test system for researchers developing new sampling algorithms.

Unlike earlier studies, we calculate the heat capacity of the peptide both *in vacuo* and in implicit solvent using the latest Amber force field. The heat capacity is particularly challenging to calculate (it is the second derivative of the partition function with respect to temperature) and provides a stringent test for sampling algorithms. This is especially the case for alanine dipeptide as the heat capacity is *almost* constant over the 800K temperature range considered and hence a large number of samples are required to reduce the statistical noise in order to clearly resolve the curve. In order to show correctness, we also calculate the heat capacity using replica exchange MD and find good agreement between the methods.

We also calculate the Ramachandran free energy surface, a common reaction co-ordinate for this system and compare the latest Amber force field results to previous theoretical and experimental alanine dipeptide studies. Finally, we discuss both the properties of GNS and peptide force fields in general.

[1] We also have to make a small number of changes to the Amber source code.

As described in the article, an alternative MD nested sampling algorithm, one which directly uses canonical MD trajectories, has also been developed, and we finish the article with a theoretical comparison between GNS, REMD and canonical nested sampling for systems which undergo a first order phase transition.

5.1 Contribution

This chapter contains a draft of a paper awaiting submission. The initial idea was formulated by Csányi. Burkoff, under the supervision of Wild and Várnai, and Baldock, under the supervision of Csányi, jointly developed the algorithm. The results in this paper were generated using an implementation of the algorithm by Burkoff, who also performed the post processing analysis included in the Results section. The initial draft of the paper was written by Burkoff with Wild and Várnai contributing to the revisions of the current draft of the paper.

Exploiting Molecular Dynamics in Nested Sampling Simulations of Small Peptides

Nikolas S. Burkoff*†, Robert Baldock*‡, Csilla Várnai†,
David L. Wild†and Gábor Csányi§

Abstract

Nested sampling (NS) is a Bayesian sampling algorithm which has significant theoretical advantages for sampling atomistic systems, particularly those which undergo a first order phase transition. It has previously been used to explore the potential energy surface of a coarse-grained protein model and has significantly outperformed parallel tempering when calculating heat capacity curves of Lennard Jones clusters. The original algorithm is Monte Carlo (MC) in nature; however, a variant, Galilean NS, has recently been introduced which allows NS to be incorporated into a molecular dynamics framework, so NS can be used for systems which lack efficient prescribed MC moves. In this work we demonstrate the applicability of Galilean NS to atomistic systems. We present an implementation of Galilean NS using the Amber molecular dynamics package and demonstrate its viability by sampling alanine dipeptide, both in vacuo and implicit solvent. Unlike previous studies, we present heat capacity curves of alanine dipeptide, whose calculation provides a stringent test for sampling algorithms. We also compare our results with those calculated using replica exchange molecular dynamics (REMD) and find good agreement. We show the computational effort required for accurate heat capacity estimation for small peptides. We also calculate the alanine dipeptide Ramachandran free energy surface and use it to compare the latest Amber force field with previous theoretical and experimental results. Finally, we discuss the behaviour of Galilean NS, comparing it to REMD and an alternative canonical molecular dynamics NS algorithm recently developed, particularly focusing on how the choice of ensembles used affects systems, such as large proteins, which undergo a first order phase transition.

Introduction

It has been over 50 years since Ramachandran and coworkers first modelled protein peptide bonds (1). In their work they used small peptides, containing only one or two peptide bonds, to study the sterically allowed protein dihedral angles. Using this information they developed the ‘Ramachandran plot’, familiar to protein scientists to this day.

The peptide bond is the smallest building block of proteins, and over the last few decades, it has continued to be studied intensively both experimentally (2–6) and theoretically (7–11). Polypeptide models and force fields of varying levels of complexity have been developed, ranging from simple coarse-grained models (12), through all-atom molecular mechanics force fields (13, 14) and hybrid quantum mechanics molecular mechanics (QM-MM) models (15, 16), to the full quantum mechanical

*These authors contributed equally to this work
†Systems Biology Centre, Senate House, University of Warwick
‡Cavendish Laboratory, University of Cambridge
§Engineering Laboratory, University of Cambridge

treatment (17). These models have allowed the computational study of peptide thermodynamics and the exploration of their potential (and free) energy surfaces (10, 11, 18, 19).

Although short peptides which occur naturally, such as the five residue neurotransmitter Met-enkephalin (20), are of particular interest, the peptide bonds in short peptides are thought to have similar properties to the peptide bonds in unfolded and unstructured proteins (21), and so their study can also inform our knowledge of proteins in their unfolded state. Peptide models have also been used to study peptide aggregation (22) and have been used in order to develop (23–25) and test (21) more general protein force field parameters and models.

Running in parallel to the development of these models and force fields, there has been considerable work in developing sampling algorithms in order to fully explore the potential (and free) energy surfaces of proteins and peptides, and to calculate accurate thermodynamics of the force fields used. These algorithms are required, as standard molecular dynamics (MD) canonical trajectories struggle to overcome energy barriers and fully sample the conformational space. As shown below, when initially developing a sampling algorithm, it is typically first tested on small peptides.

The *de facto* standard algorithm for general conformational phase space exploration is replica exchange molecular dynamics (REMD) (26). A set of canonical MD trajectories are run with each 'replica' using a different temperature. Periodically, the swapping of conformations for two replicas is proposed and is accepted using the standard Metropolis-Hastings Monte Carlo (MC) acceptance criteria. The high temperature replicas ensure the system can escape from local modes. Many extensions, such as allowing the temperature of the replicas to change throughout the simulation in order to improve efficiency, have been developed (27). In the original REMD research, the penta-peptide Met-enkephalin was studied (26), and subsequently the method has been very widely used, for example, to fold the Trp-cage mini-protein (28) and to calculate the heat capacity curve of an SH3 domain (29).

One of the main thermodynamical properties of interest to protein scientists is the free energy difference between different states of the system. These are used to plot the free energy surface with respect to reaction co-ordinates of interest and give key insights into the behaviour of the system. Although in this work we focus on algorithms which do not require suitable reaction co-ordinates to be known *a priori*, if these are available, then specialized free energy algorithms can be used to calculate such differences (30–32). One such algorithm is umbrella sampling (30), where an extra force is applied to keep the reaction co-ordinate at a chosen value. Originally tested on Lennard Jones (LJ) clusters, umbrella sampling has been used to study short peptides (33) and is now a standard free energy calculation algorithm.

Sophisticated general (i.e. not requiring prior knowledge of the potential energy surface) conformational sampling algorithms have also been developed. For example, accelerated molecular dynamics, where a bias (a function of *only* the potential energy) is used to facilitate the traversal of energy barriers (34). To initially test the algorithm, in the original work, Hamelberg *et al.* calculate the free energy surface of alanine dipeptide, a simple molecule with only a single peptide bond (34). Another example is multicanonical sampling, using either Monte Carlo (35) or molecular dynamics (36) sampling. In this algorithm, instead of sampling from the Boltzmann distribution – $\mathbb{P}(\Omega) \propto \exp(-E(\Omega)\beta)$ – samples are drawn from the *multicanonical* distribution – $\mathbb{P}(\Omega) \propto 1/g(E(\Omega))$, where $g(E)$ is the density of states. Multicanonical sampling was specifically designed to be efficient when sampling systems which undergo a first order phase transition (35). Multicanonical MD has been used to study the free energy landscapes of tri-peptides (37) and a seven-residue DNA binding peptide (38). Recently, the algorithm has been applied to larger peptides and protein domains; further applications of the multicanonical MD algorithm can be found in a recent review (39). Many variants of the multicanonical algorithm, such as the Wang-Landau algorithm (40), have also been developed. Further examples of MC algorithms include equi-energy (41) and well-tempered ensemble (42) sampling.

Recently, Skilling introduced a novel sampling technique, nested sampling (43), which has distinct advantages for sampling atomistic systems, and subsequently, a similar algorithm to nested sampling, the energy partitioning method, was independently developed for sampling water molecules and binary mixtures of fluids (44, 45).

Nested Sampling

Nested sampling is a Bayesian sampling algorithm, specifically designed to sample high dimensional spaces (43, 46). The algorithm is designed for systems where the bulk of the probability mass is contained in an exponentially small volume of phase space. The algorithm outputs both an estimate for the evidence (also known as the marginal likelihood or partition function) and a set of samples from which thermodynamic variables, such as heat capacities and free energy differences, can be calculated at any temperature.

Whilst initially developed for statistical inference (43), the algorithm is well-established in the astrophysics community (47) and has also been successfully applied in a variety of other fields including bioinformatics (48), systems biology (49) and flow model selection (50).

The nested sampling algorithm has also been applied to atomic systems. Pártay *et al.* have used it to sample LJ clusters (51) and hard sphere models (52), and in our previous work we explored the potential energy surface of a coarse-grained protein model (53). The algorithm is ideally suited for the sampling of atomistic systems, and particularly proteins as, for these systems, the dimension of the phase space is large; accessible conformations (at temperatures of interest) are located in exponentially small regions of phase space; and, although the nested sampling procedure is *athermal*, a single nested sampling simulation can be used to estimate thermodynamic observables at any temperature.

Furthermore, nested sampling has also been shown to be particularly efficient at sampling systems which undergo a first order phase transition (46). For example, when calculating the heat capacities of LJ clusters, nested sampling outperformed parallel tempering by over an order of magnitude (51). It is well known that proteins undergo a first order phase transition upon thermal unfolding (54–56).

The original nested sampling algorithm is a Monte Carlo (MC) sampling algorithm and in our previous work, the coarse-grained protein model used was designed to allow efficient MC crankshaft moves (53). For example, all bond lengths were fixed and the peptide bond was kept exactly planar. In this work, however, we apply the algorithm to an all-atom force field where the extra degrees of freedom would make MC sampling more challenging.

Recently, however, Skilling introduced *Galilean* nested sampling (57), a variant of the nested sampling algorithm for which each atom is given velocities and system specific MC moves are not required. The velocities are then used to evolve sample points using Galilean sampling, a novel exploration procedure, rather than using the standard Hamiltonian or canonical exploration.

Galilean exploration allows nested sampling to be used for systems for which required MC moves are not efficient. For example, all-atom models of proteins and peptides, especially with explicit solvent, have a large number of degrees of freedom and MC sampling requires moves such as bond stretching and angle bending. These moves are not particularly efficient, leading to long decorrelation times, especially at low temperatures. It also allows an efficient nested sampling algorithm to be implemented within a molecular dynamics (MD) framework and, hence, increases the utility of the nested sampling method.

In this work we implement Galilean nested sampling within the Amber MD package (13) and we test the algorithm by generating thermodynamical data for the short peptide alanine dipeptide, both *in vacuo* and in implicit solvent. Unlike earlier work with alanine dipeptide, we focus on calculating accurate[1] heat capacity curves and compare the nested sampling results to those obtained using the

[1] In this work, 'accurate' thermodynamic data indicates that it is accurate *for the force field used.*

standard REMD procedure. We also calculate dihedral angle Ramachandran free energy surfaces, comparing the results to previous theoretical and experimental work.

We discuss the properties of Galilean nested sampling and our expectations for the method looking to the future. Finally, Nielsen has recently published an alternative MD version of nested sampling, which uses canonical MD exploration (58). We end the discussion with a description of the differences between this algorithm, REMD and Galilean nested sampling, with a particular focus on their behaviour when sampling systems which undergo a first order phase transition.

Methods

Following the principles of classical statistical mechanics, the configurations of constant volume systems which are in thermal equilibrium with their surroundings are distributed according to the Boltzmann (or canonical) distribution. Specifically, at temperature T, the probability of the system adopting the configuration Ω is proportional to $\exp(-E(\Omega)\beta)$ where $\beta = 1/(k_B T)$, k_B is the Boltzmann constant ($\approx 2\text{x}10^{-3}$ kcal/mol/K) and $E(\Omega)$ is the potential energy of configuration Ω.

The normalization constant of the Boltzmann distribution, the partition function,

$$Z(\beta) = \int_{\Omega} \exp(-E(\Omega)\beta) d\Omega,$$

is of fundamental importance in statistical physics, as it can be used to obtain thermodynamic quantities; for example, the internal energy,

$$U = \langle E(\Omega) \rangle_{\beta} \equiv - \left(\frac{\partial \ln Z}{\partial \beta} \right)_V ,$$

and the constant volume configurational heat capacity,

$$C_v = \langle E^2(\Omega) \rangle_{\beta} - (\langle E(\Omega) \rangle_{\beta})^2 \equiv k_B \beta^2 \left(\frac{\partial^2 \ln Z}{\partial \beta^2} \right)_V ,$$

where $\langle . \rangle_{\beta}$ is expectation under the Boltzmann distribution.

Although it is possible to estimate the partition function using the 'harmonic mean approximation', $Z^{-1} = \langle \exp(E(\Omega)\beta) \rangle$, this estimate has infinite variance and hence should be avoided (59).

Nested Sampling Algorithm

The nested sampling algorithm is an iterative procedure which generates a set of energy levels $E_1 > E_2 > E_3 \ldots$, where for each i, E_i is chosen so that

$$\frac{\int_{\Omega} \mathbb{I}\{E(\Omega) < E_i\} d\Omega}{\int_{\Omega} \mathbb{I}\{E(\Omega) < E_{i-1}\} d\Omega} = \alpha$$

for some fixed proportion α. Hence the algorithm takes steps equidistant in 'the logarithm of phase space volume'. See Figure 1 for further details.

The proportion $\omega_i = \alpha^{i-1} - \alpha^i$ of conformations have energy between E_{i-1} and E_i and hence, by using numerical integration, we can estimate the partition function to be

$$Z(\beta) = \int_{\Omega} \exp(-E(\Omega)\beta) d\Omega \approx \sum_i \omega_i \exp(-E(\Omega_i)\beta). \tag{1}$$

The algorithm does not prescribe a specific terminating condition, only running until the variables

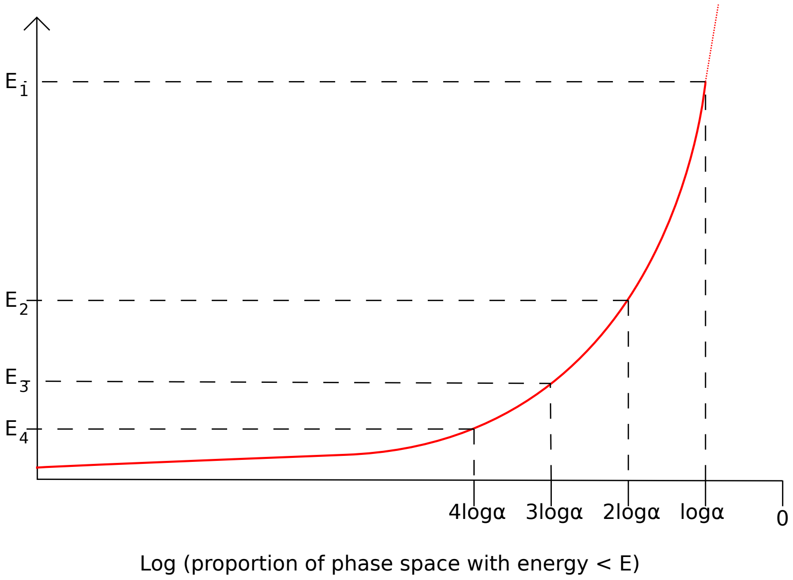


Figure 1: The energy levels $E_1 > E_2 > \ldots$ are chosen to be equidistant in log phase space volume. Therefore, the proportion $1 - \alpha$ of conformations have energy $> E_1$ and $\alpha - \alpha^2$ of conformations have energy $< E_1$ and $> E_2$.

of interest have converged. In previous work the algorithm was terminated at iteration j when

$$\log\left(\sum_{i=1}^{j}\omega_i \exp(-E(\Omega_i)\beta)\right) - \log\left(\sum_{i=1}^{j-1}\omega_i \exp(-E(\Omega_i)\beta)\right) < \epsilon$$

for the lowest temperature, $T_{\min}$, (respectively highest β) of interest (53). We follow the same procedure here, and by setting $\epsilon = 10^{-5}$, we ensure the heat capacity estimate has converged at $T_{\min}$.

Generation of Energy Levels

Although the original algorithm does not prescribe a specific method to calculate the energy levels, a Monte Carlo method is proposed and is described in Algorithm 1. An active set of K samples, uniformly distributed over the set of configurations with energy below the current energy level, is maintained. The set is initialized with samples uniformly distributed throughout the whole phase space. The energy of the highest energy configuration in the active set is chosen to be the first energy level, E_1, and this configuration, Ω_1, is removed from the active set.

The $K - 1$ samples remaining in the active set are uniformly distributed over the set of configurations with energy below the current energy level, and only a single new configuration is required. This configuration is generated by copying an existing member of the active set and taking the final configuration of a Markov chain with equilibrium distribution given by $\mathbb{P}(\Omega) \propto \mathbb{I}\{E(\Omega) < E_1\}$. The second energy level E_2 is then taken to be the energy of the highest energy configuration currently in the active set, Ω_2, and the procedure repeats, generating $E_3, E_4, \ldots$.

At iteration i, the proportion of the configuration space with energy $< E_{i-1}$ which has energy less than the sample of the active set with highest energy $\sim \mathrm{Beta}(K+1, 1)$ and its expected value is $K/(K+1)$. Therefore, ignoring uncertainties, we find $\alpha = K/(K+1)$. It is possible to quantify these uncertainties when producing estimates of the partition function (46).

The sample points removed from the active set, $\{\Omega_1, \Omega_2, \ldots\}$ can be used to estimate properties of the Boltzmann distribution at any temperature. Ω_i represents ω_i of configuration space, and therefore represents $\chi_i(\beta) = \omega_i \exp(-E(\Omega)\beta)/Z(\beta)$ of the probability mass of the Boltzmann distribution at

Algorithm 1 Monte Carlo (MC) algorithm to generate nested sampling energy levels

```
Generate $K$ samples uniformly distributed throughout phase space, the active set
$i \leftarrow 1$
loop
    Remove sample with highest energy, $E^*$, from the active set
    Output $E_i = E^*$
    Copy randomly chosen member of active set to use as a starting conformation for a MC chain
    Run Markov Chain Monte Carlo with equilibrium distribution $\propto \mathbb{I}\{E(\Omega) < E^*\}$
    Add the final conformation from MC chain into the active set
    $i \rightarrow i + 1$
end loop
```

inverse thermodynamic temperature β. Any property $Q(\Omega|\beta)$ can be estimated as

$$\mathbb{E}(Q|\beta) = \sum_i \chi_i(\beta) Q(\Omega_i).$$

For example, the heat capacity is given by

$$C_v(\beta) \approx k_B \beta^2 \left[\sum_i \chi_i(\beta) E^2(\Omega_i) - \left(\sum_i \chi_i(\beta) E(\Omega_i) \right)^2 \right]. \tag{2}$$

Estimates for (Helmholtz) free energy differences can also be computed: if the set of samples $\{\Omega\}$ can be split into disjoint macrostates A and B, then the free energy difference is given by

$$F_A - F_B \approx -\beta^{-1} \left[\log \left(\sum_{\{i:\Omega_i \in A\}} \omega_i \exp(-E(\Omega_i)\beta) \right) - \log \left(\sum_{\{i:\Omega_i \in B\}} \omega_i \exp(-E(\Omega_i)\beta) \right) \right]. \tag{3}$$

Although thermodynamic variables can be calculated for any temperature, the energy levels that are output by nested sampling and the MC exploration it uses are both athermal. In the Discussion Section we show how these imply nested sampling is ideally suited for sampling systems which undergo a first order phase transition.

Galilean Exploration

In our previous work sampling protein models we used a coarse-grained force field, CRANKITE (60). In this model each amino acid had 3 degrees of freedom, the dihedral angles ϕ and ψ and the C_α valence angle, and we used crankshaft rotations as MC moves which efficiently sample the configurational space (53). However, more realistic all-atom models have more degrees of freedom, and in order to sample the system, additional MC moves such as angle bending and bond stretching, must be included. These moves, especially at low temperatures, or for systems which include explicit solvent molecules, are often inefficient. For these systems, sampling using MD, which has shorter decorrelation times than MC, is often preferred.

In this work we implement *Galilean* exploration, a method of exploration used to generate nested sampling energy levels, recently introduced by Skilling (57). Galilean exploration does not require system-specific MC moves. Following the MD approach, the atoms of the conformation are given velocities and the system is then evolved along a trajectory generating samples uniformly distributed over all conformations with energy less than a prescribed value. The details of Galilean exploration are given below.

In Galilean nested sampling, in order to generate a new sample for the active set, an existing member of the set, Ω, with atomic co-ordinates $\mathbf{x}$ is chosen. A set of velocities, $\mathbf{v} : v_i \sim \mathcal{N}(0, k_B T)$, for a chosen parameter,[2] T, are drawn and the move $\mathbf{x} \to \mathbf{x}' = \mathbf{x} + \tau\mathbf{v}$ is proposed, where τ is the timestep.[3] If the proposed conformation has energy below the current energy level, the move is accepted; otherwise we try to 'reflect' the conformation back into the acceptable region by choosing a unit normal vector $\mathbf{n}$ and proposing the move $\mathbf{x} \to \mathbf{x}'' = \mathbf{x}' + \tau(\mathbf{v} - 2\mathbf{n}(\mathbf{n}.\mathbf{v}))$. In principle any unit vector $\mathbf{n}$ can be used. However, if possible, we would like to reflect off the boundary of the acceptable region, thus ensuring the move is accepted. We can estimate this orientation by taking $\mathbf{n}$ as the unit vector in the direction of $\nabla E(\mathbf{x}')$.

If the new conformation has energy less than the current energy level, the reflection is accepted and the trajectory continues with velocity $\mathbf{v}' = \mathbf{v} - 2\mathbf{n}(\mathbf{n}.\mathbf{v})$. If not, detailed balance insists we reject the move, thus remaining at $\mathbf{x}$, and we continue the trajectory by using the velocity $-\mathbf{v}$. See Figure 2 for an example trajectory.

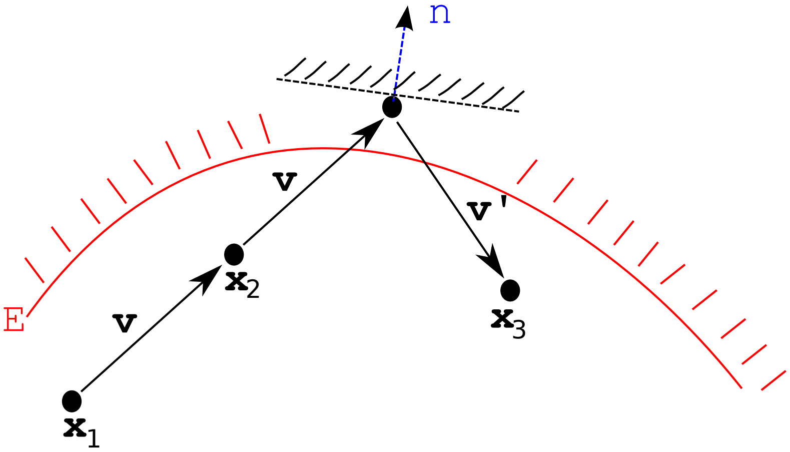


Figure 2: At $\mathbf{x_1}$ the conformation $\mathbf{x_2} = \mathbf{x_1} + \tau\mathbf{v}$ is proposed. As the conformation remains in the acceptable region (has energy below E, shown by the red contour), it is accepted. The proposed move to $\mathbf{x}' = \mathbf{x_2} + \tau\mathbf{v}$ takes the conformation outside the acceptable region, so it is reflected to $\mathbf{x_3} = \mathbf{x}' + \tau\mathbf{v}'$, where $\mathbf{v}' = \mathbf{v} - 2\mathbf{n}(\mathbf{n}.\mathbf{v})$ for the unit vector $\mathbf{n} = \nabla E(\mathbf{x}')$. As $\mathbf{x_3}$ is inside the acceptable region, the move is accepted and the trajectory continues using velocity $\mathbf{v}'$. If $\mathbf{x_3}$ were to be outside the acceptable region, then the move would have been rejected, the conformation returned to $\mathbf{x_2}$ and the velocity reversed to $-\mathbf{v}$.

Unless the energy level boundary is crossed, the same $\mathbf{v}$ continues to be used throughout the trajectory, as the induced systematic motions can be expected to explore more efficiently than random diffusions. However, in order to decrease equilibration time, it is suggested that we slightly perturb the velocity at each iteration and, instead of using velocity $\mathbf{v}$, use the velocity $\mathbf{v^p} = \mathbf{v}\cos\theta + \tilde{\mathbf{v}}\sin\theta$ where $\tilde{\mathbf{v}}$, is a newly drawn set of velocities and θ is small.

Galilean Nested Sampling for Peptides

In this work we adapt the Amber molecular dynamics package (13) to perform Galilean exploration in order to generate nested sampling energy levels. We use the Amber ff12SB protein force field with `igb=6` for vacuum simulations and, for implicit solvent simulations, `igb=8`, a generalized Born solvation model (61). We use the default Amber chirality and *trans/cis* peptide bond restraints. We do not employ a van der Waals distance cutoff and do not constrain the covalent hydrogen bond distances with SHAKE.

[2] The parameter T controls how fast the particle moves and hence is analogous to temperature in canonical MD simulations. However, it does not correspond to the temperature of any canonical MD simulation.

[3] With Galilean exploration, there is a direct correspondence between timestep τ and 'temperature' T: the transformation $(T, \tau) \to (aT, \tau/\sqrt{a})$ for constant a is invariant. In this work, for each simulation, we fix τ and allow T to vary as described later in the text.

Due to the rapid shrinking of the available phase space volume, we find it is sufficient to set $\alpha = 0.5$, and thus remove half of the available phase space each iteration. However, it is necessary to estimate the next energy level to a very high degree of accuracy, and following the standard algorithm, which would use an active set with a single sample, is inappropriate. Therefore, instead, at each iteration we use Galilean exploration to generate a large set of uniformly distributed samples and use the empirical median of this distribution to estimate the next energy level. The samples from this iteration with energy less than the next energy level are still uniformly distributed, so we can, as in the original algorithm, re-use these samples in subsequent iterations. The starting conformations for the trajectories of the subsequent iteration are chosen uniformly from the set of conformations with energy less than the new energy level.

As the accessible region of phase space shrinks, it is necessary to reduce the magnitude of the velocities in order to keep the trajectories within the allowed region. We define the *mean free path* to be the average number of successful steps taken before requiring a reflection. We use the variable T in order to keep the mean free path constant throughout a nested sampling simulation. See the Supporting Material for technical details.

In the original algorithm, simulations are initialized by choosing samples uniformly throughout the whole of configuration space. As we are only interested in thermodynamics at relatively low temperatures, we initialize the algorithm by generating a set of samples uniformly distributed over the conformations with potential energy below a chosen initial energy level. We refer the reader to the Supporting Material for further details concerning initializing the algorithm at a specific energy level.

Each reflection requires two separate force evaluations, one when the sample steps outside the acceptable region and one after it has been reflected. Therefore, when the mean free path is lower, there are more reflections, and so trajectories must be shortened in order to maintain the same number of force evaluations,[4] and therefore computational expense, when comparing efficiencies. Due to the implementation within Amber, in this work we calculate the forces at each step of the trajectory. However, it is important to note that this is not strictly required as Galilean exploration only requires the calculation of the forces (i.e. $-\nabla E$) when outside the acceptable region. At other times, only the potential energy is required (to check whether the trajectory has left the acceptable region).

Results

We demonstrate the Galilean nested sampling algorithm by using it to calculate the thermodynamics and free energy surfaces of the small peptide alanine dipeptide both in vacuum and implicit solvent.

Alanine dipeptide in vacuo

It is over 50 years since Ramachandran and co-workers analysed the sterically allowed peptide dihedral angles ϕ and ψ. In their work they introduced the name dipeptide to describe molecules which include, beside a single amino acid, adjacent residues as far as the C_α atoms (1). Over the last 50 years dipeptides, and particularly alanine dipeptide (N-acetyl-alanyl-N'-methylamide; see Figure 4), have been studied experimentally, both in solution (6) and in the gas phase (4). Alanine dipeptide has also been studied from a quantum mechanical perspective (8, 9) and has previously been used to parametrize molecular force fields (62) and test their accuracy (21).

Alanine dipeptide, alongside glycine dipeptide, is the smallest molecule which contains a peptide bond and therefore, due to its size, the molecule is commonly used when testing new sampling

[4] Specifically, the total number of force evaluations = $(m + 2)S/(m + 1)$, where S is number of steps and m the mean free path.

algorithms, for example (11). For the same reason we have chosen to use alanine dipeptide to test Galilean nested sampling. Unlike previous work, here we focus not only on calculating the free energy (or potential energy) surface, but also on the accurate determination of the heat capacity of the system. To the authors' knowledge these are the first published heat capacity curves for alanine dipeptide.

Heat Capacity

Figure 3 shows estimates for the heat capacity of alanine dipeptide *in vacuo* for five independent nested sampling simulations (lines) calculated using Equation (2). Although the potential energy at temperatures of interest is low (e.g. at 360K, $U \approx 0 \pm 4$ kcal/mol), the initial energy cutoff was chosen to be $E = 100$ kcal/mol. This is necessary due to the extremely high energy barrier separating room-temperature accessible conformations with dihedral angle $\phi > 0$ and those with $\phi < 0$ (see Figures 4 and 5 below). Although for biophysical systems we would not normally be interested in the behaviour of the system at 100K, for this study, we choose $T_{\text{min}} = 100$ as this allows us to capture the peak in the heat capacity curve.

Following our previous work (53), we choose to use a large number of independent walkers, in this case 16000. We use the parameter T to keep the mean free path ≈ 2, and by setting $\theta = 0.2$ we allow a small amount of velocity randomization every Galilean step. See the Discussion Section for further details concerning the chosen parameters.

Each Galilean trajectory runs for a total of 2700 steps, outputting the potential energy every 75 steps. This implies each nested sampling iteration uses approximately 57 million force evaluations, which leads to a total of $\approx 9.6\text{x}10^9$ force evaluations per simulation. This is a *very* large number of force evaluations for such a small system. However, as the value of the heat capacity only varies by $\sim 2k_B$ over the 800K temperature range, a very large number of force evaluations are required to reduce the statistical error to a small enough value to clearly resolve the curve.

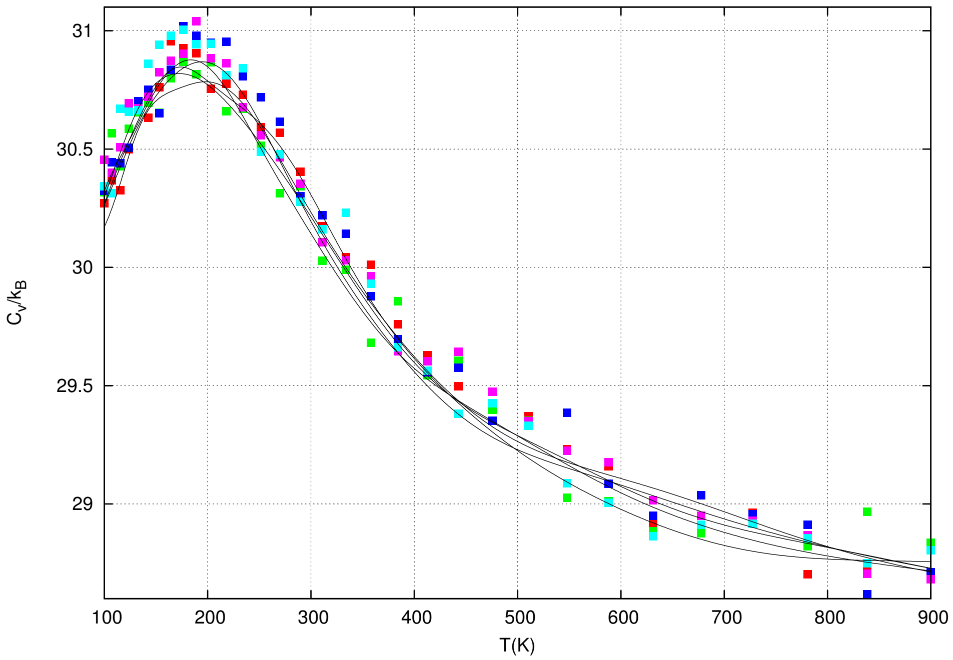


Figure 3: The heat capacity, C_v, from 5 independent nested sampling simulations (lines) and 5 REMD simulations (points). All simulations used a comparable number of force evaluations ($\approx 9.6\text{x}10^9$). See the text for further details.

The variance between estimates from independent simulations is very small. However, in order to show that the algorithm has converged to the *correct* value, Figure 3 also includes heat capacity estimates from five independent REMD simulations, and there is good agreement between the methods. The REMD simulations use a similar number of total force evaluations (9.6x10^9) as the nested sampling simulations. The temperatures of the 32 replicas are in a geometric progression from 100 to 900K, and swaps between different replicas are attempted every 2ps. A very small timestep (0.2fs) was used to ensure accuracy, and the parameters of these REMD simulations (such as the number of replicas) have not been optimized. Therefore, we claim only that nested sampling and REMD are of similar efficiencies for this system. Rigorous benchmarking of REMD and Galilean nested sampling on larger systems is the major focus of our current work.

Free Energy Surface

A standard free energy reaction co-ordinate for alanine dipeptide is the pair of dihedral angles (ϕ, ψ), for example see (11). We split the conformations generated by a nested sampling simulation into separate 'bins' based on their dihedral angles and then use Equation (3) to generate the free energy surface. A Gaussian filter has then been applied to smooth the data and the result, for 300K, is shown in Figure 4. For comparison, the unsmoothed free energy surface is shown in the Supporting Material; see Figure S3. When using the original nested sampling algorithm, each energy level corresponded to exactly one sample point, which represented ω_i of phase space. In this work, we output a whole set of samples for each energy level and, when calculating free energy surfaces, we give each sample a uniform[5] share of the weighting ω_i.

Although the focus of this work is the implementation of Galilean nested sampling rather than force field development, it is nevertheless interesting, to use these results to compare the Amber ff12SB force field with experimental results and quantum mechanical (QM) calculations. Pohl *et al.* compared alanine dipeptide QM calculations with infra-red absorption spectra in Ar and Kr isolation matrices (4). From QM calculations they found that the two most common conformations were expected to be $C_{7\mathrm{eq}}$ (also named γ_{L}) and C_5 ($\beta_{\mathrm{L(D)}}$). Depending on the choice of basis sets, the relative abundance of $C_{7\mathrm{eq}}$ (at 343K) was between 32% and 63%. For this force field, we also find the same two common conformations with the abundance of $C_{7\mathrm{eq}}$ (at 343K) $\approx$ 66%. These conformations were also identified experimentally (4). Tobias *et al.* compared a QM and molecular mechanics (MM) potential energy surface of alanine dipeptide and, although they find differences in the position of local minima, they conclude that the MM force field provides a very good description of alanine dipeptide *in vacuo*. We find the locations of minima agree well with the positions on the MM force field used by Tobias *et al.*.

These nested sampling simulations converge the heat capacity, which requires the calculation of the second derivative of the logarithm of the partition function (as a function of temperature). Free energies are calculated directly from the logarithm of the partition function, without differentiation, and therefore it is unsurprising that for the simulations from Figure 3 we find excellent agreement when calculating free energies; for example, see Figure S3 in the Supporting Material. If the heat capacity were not required, only the free energy surface, it is likely that much smaller simulations could be used. In this case, fewer force evaluations would be needed.

For alanine dipeptide, there is a clear choice of reaction co-ordinates for a low dimensional free energy surface (the dihedral angles), and, as the system is so small, Figure 4 could easily be calculated by a specialized free energy calculation method such as umbrella sampling (30). However, these methods typically require a reaction co-ordinate to be chosen *a priori*. This is not the case for

[5] Technically, as samples have slightly different energies ($E_i > E(\Omega) > E_{i+1}$), they ought to represent slightly different proportions of phase space. However, the energy gap between successive energy levels is extremely small, and this approximation is analogous to the approximation used by the original nested sampling algorithm when performing the numerical integration to estimate the partition function. Therefore, in practice, we find this approximation adequate.

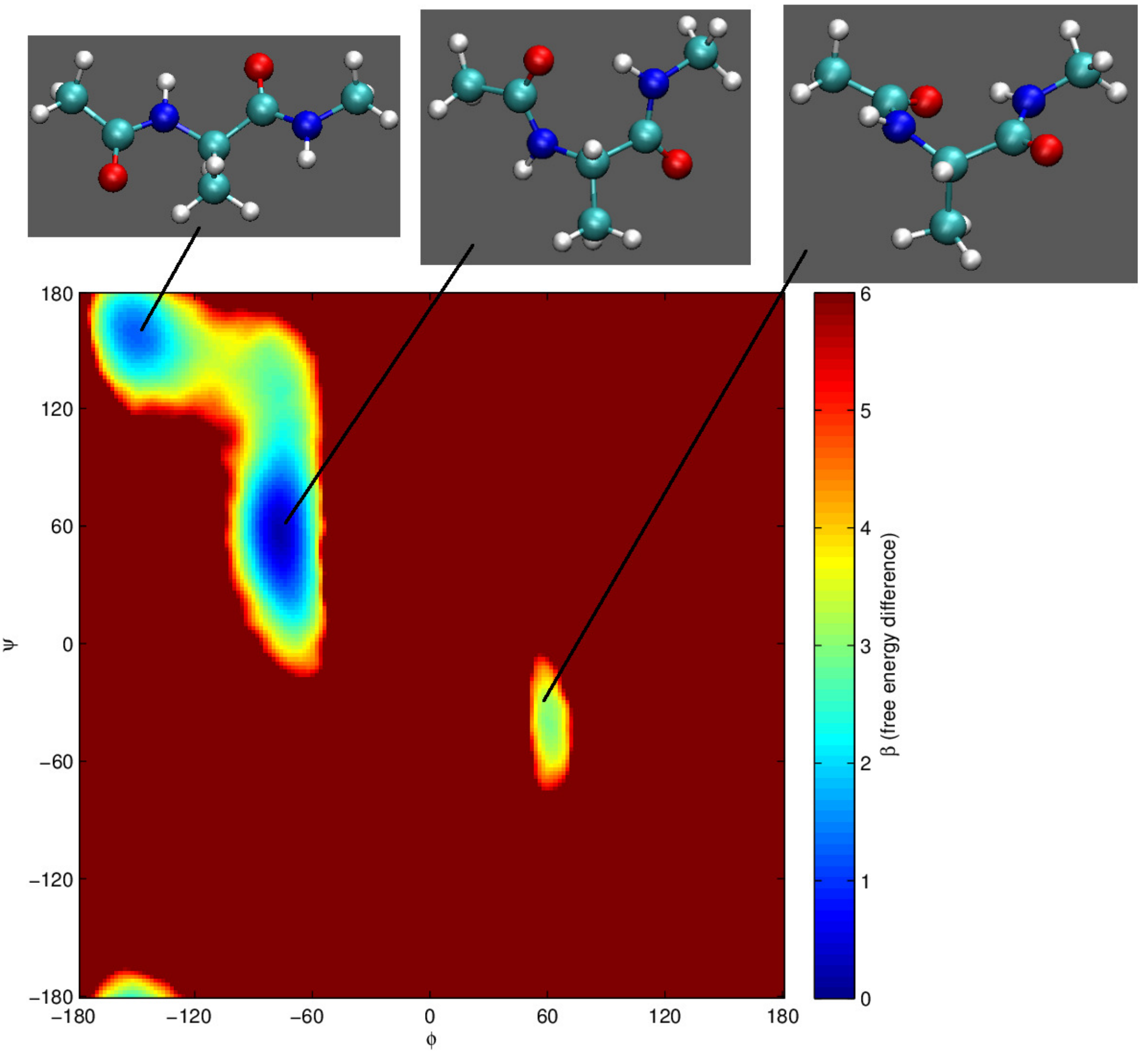


Figure 4: Top: Three conformations of alanine dipeptide accessible (*in vacuo*) at room temperature, from left to right C_5, $C_{7\mathrm{eq}}$ and $C_{7\mathrm{ax}}$; see (4, 8, 9). Bottom: The free energy surface of alanine dipeptide in vacuuo at 300K. See the text for further details. Note: in this work, the dark red used for β(free energy difference)= 6 is also used where this value is greater than 6.

nested sampling as no reaction co-ordinate is required for the sampling algorithm. A discrete order parameter can actually be derived directly from clustering the samples output by a nested sampling simulation; we refer the reader to (51) for further details.

By reweighting the samples from the same nested sampling simulation, the free energy surface can be calculated for arbitrary temperatures. For example, Figure 5 shows the free energy surface at 100, 200 and 900K. Although there is a clear energy barrier at $\phi \approx 0$, it is possible for canonical trajectories at 900K to overcome this barrier. However, at 600K it is all but impossible. This shows the importance of ensuring there are replicas which have temperatures high enough to overcome all energy barriers when running REMD. If not, truly accurate thermodynamics of a system cannot be attained. Further discussion concerning this can be found in the Supporting Material.

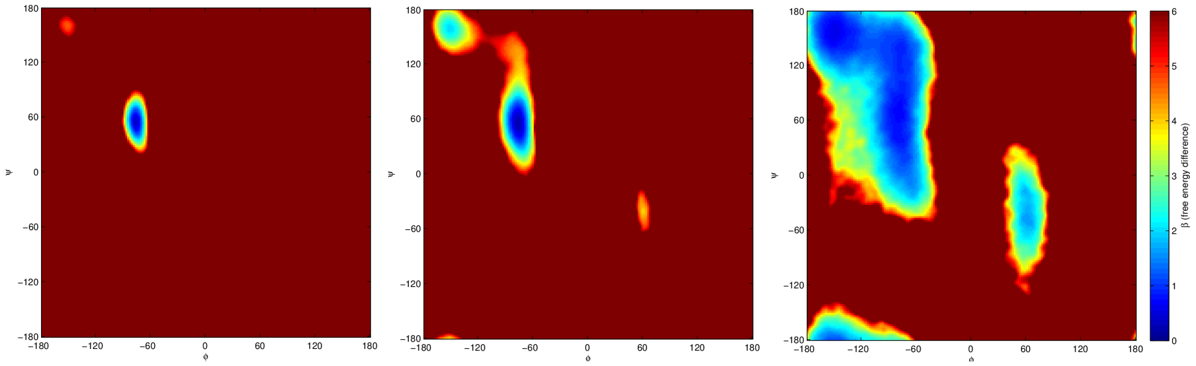


Figure 5: Free energy surface of alanine dipeptide *in vacuo* at 100K, 200K and 900K.

Alanine dipeptide in implicit solvent

In this section we perform nested sampling of alanine dipeptide in solvent and compare the results generated by the Amber ff12SB force field to the latest experimental data.

There is no theoretical barrier to using Galilean nested sampling algorithm with explicit solvent molecules, as each solvent molecule can be given velocities and the whole system can be evolved using Galilean exploration. However, in this work we have focused on the calculation of accurate heat capacity curves and so, in order to reduce computational expense, we have chosen to use a generalized Born (61) implicit solvent model.

The initial energy level was chosen to be 75 kcal/mol, which is high enough to allow the heat capacity to be calculated at 900K, similarly to the *in vacuo* case. All other parameters have been kept the same, except that in order to capture the peak of the heat capacity curve, we set $T_{\mathrm{min}} = 30$. Therefore we needed to calculate an additional 48 energy levels and, as we chose to use the same number of force evaluations for the simulations as previously, these additional iterations meant we had to shorten trajectory lengths from 2700 to 2100 steps.

Heat Capacity

Figure 6 shows the heat capacity of alanine dipeptide in implicit solvent. There is, again, good agreement between nested sampling and REMD simulations. In this case, the 32 temperatures of the REMD replicas were chosen in geometric progression from 30 to 900K. It is interesting to note the peak of the curve is ≈ 140K lower than the *in vacuo* case.

Free Energy Surface

Analogous to the *in vacuo* case, the dihedral angle free energy surface of alanine dipeptide in solvent can be calculated using the samples output from a nested sampling simulation. Figure 7 shows the free energy surface at 300K together with images of the three low energy minima, P_{II}, β and α_{R}, as defined by (6).

The results presented here clearly show that there are three free energy minima, and their locations are given by P_{II} $(-80^\circ, 150^\circ)$, β $(-150^\circ, 150^\circ)$ and α_{R} $(-75^\circ, -20^\circ)$. These results agree qualitatively with those from a published QM/MM force field ((21) Figure 6).

It is possible to directly compare the accuracy of this implicit molecular force field with experimental results. Each conformation from nested sampling is assigned to a 'basin', P_{II}, β, α_{R} or 'other'; the basins are defined in Figure 8. The choice for basin definitions has been guided by the free energy surface, rather than previous definitions found in the literature. However, the occupancy probabilities shown in Figure 9 are not sensitive to the precise definitions used. Using Equation (3), free energy differences, and hence probabilities of occupancy (i.e. $\mathbb{P}(\Omega \in P_{\mathrm{II}}|T)$, with T the canonical

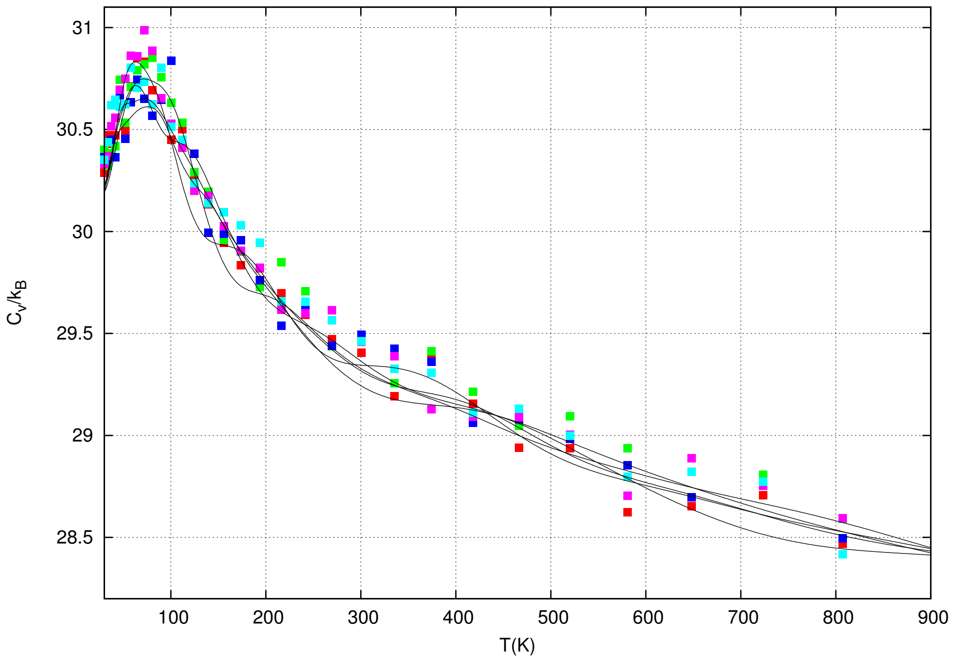


Figure 6: The heat capacity, C_v, of alanine dipeptide in implicit solvent from 5 independent nested sampling simulations (lines) and 5 REMD simulations (points). All simulations used a comparable number of force evaluations ($\approx 9.6\text{x}10^9$). See the text for further details.

temperature), can be calculated. Figure 9 compares these probabilities of occupancy with probabilities derived from published ATR-absorbance spectra data (6). The experimental results are shown by squares and the estimates calculated from the nested sampling simulations are shown by the error bars (mean±sd of 5 independent simulations). The nested sampling probabilities of occupancy for 'other' ($\approx 2\%$) are not displayed.

Figure 9 shows that this protein force field (ff12SB), together with the generalized Born implicit solvent `igb=8`, overestimates the probability of finding the molecule in the α_R conformation. The inaccuracy of peptide force fields is discussed in the following section.

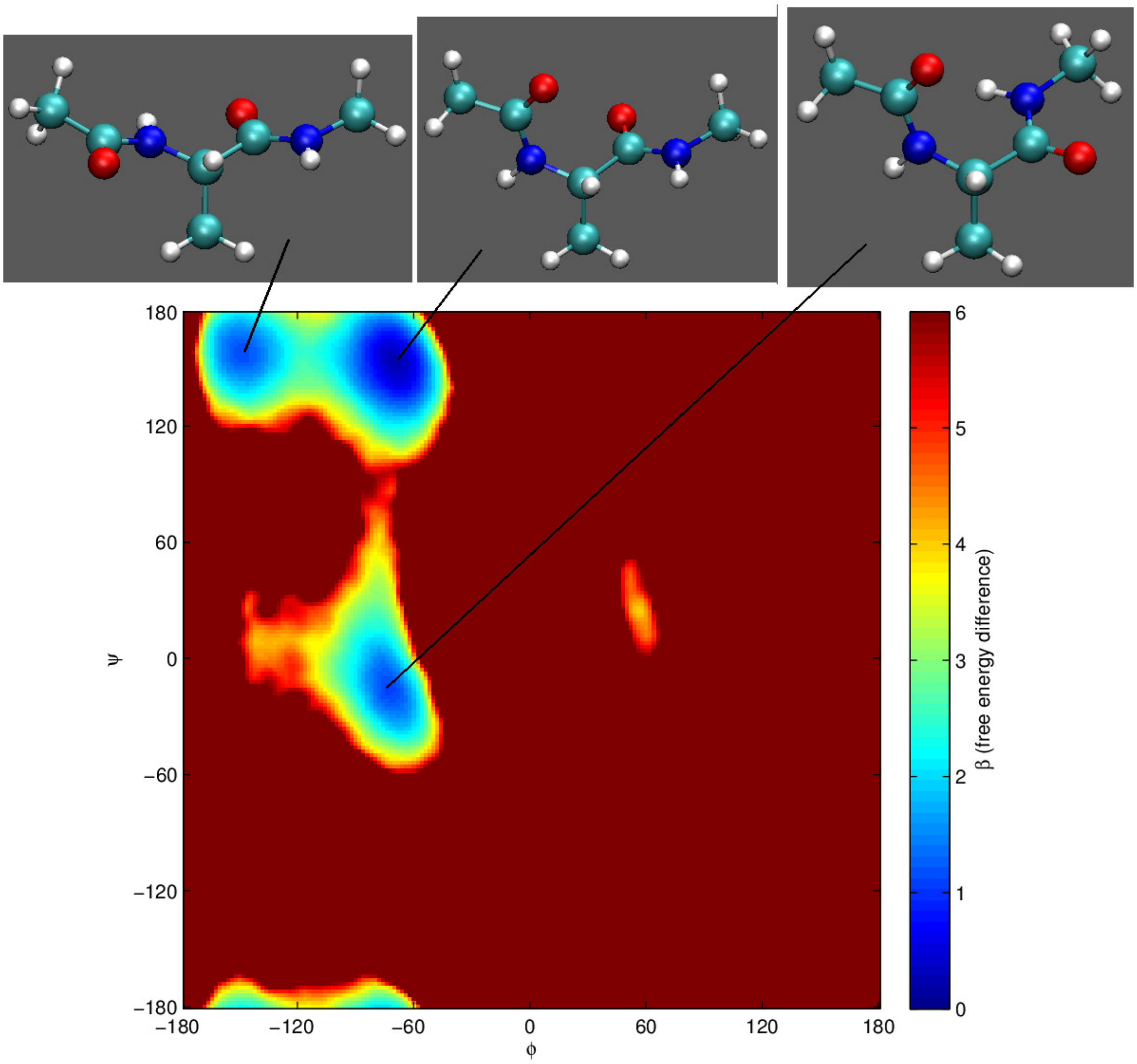


Figure 7: Bottom: The Ramachandran free energy surface at 300K for alanine dipeptide. Top: Conformations from the three main accessible regions of the Ramachandran plot, from left to right β, P_{II} and α_{R} (6)

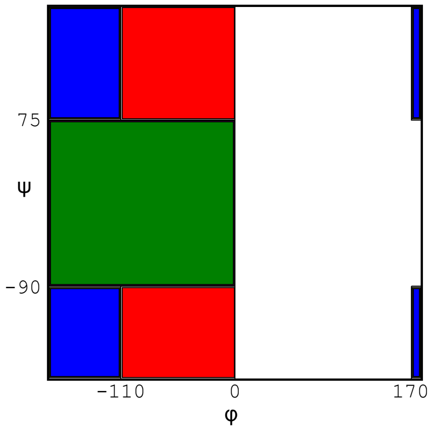


Figure 8: The dihedral angle definitions of P_{II} (red), β (blue), α_{R} (green) and 'other' (white). The choice for basin definitions has been guided by the free energy surface, rather than previous definitions found in the literature; however, the occupancy probabilities shown in Figure 9 are not sensitive to the precise definitions used.

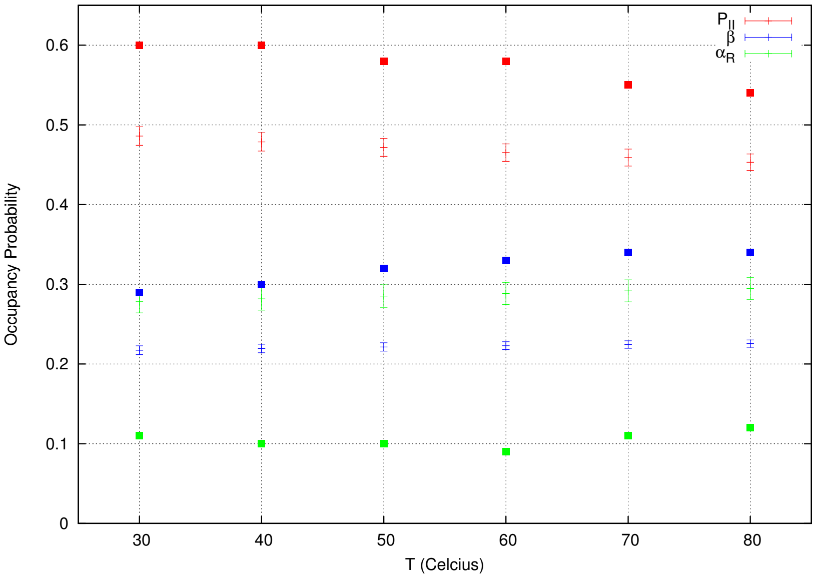


Figure 9: The occupancy probabilities for the three main conformations P_{II} (red), β (blue) and α_{R} (green), as a function of temperature. The squares are ATR-absorbance spectra data (6); the error bars are mean±sd of 5 independent nested sampling simulations. The nested sampling probabilities of occupancy for 'other' ($\approx 2\%$) are not displayed. The 'other' refer to the small population of α_{L} (left-handed helical) conformations, the free energy minimum with $\phi > 0$ on Figure 7. No direct test to detect these conformations was possible using the experimental techniques used in (6).

Discussion

Unlike proteins where in general, there is a dominant free energy minimum (the native state), the free energy differences between the different macrostates in peptides is small, and hence a distribution of states exist when the peptide is in thermal equilibrium. Therefore, a small inaccuracy in a protein force field can effect a large change in the equilibrium distribution when compared to experiments. We find that this is the case with the Amber force field used here, with the α_{R} conformation being over-represented.

The Amber force field was originally developed to study proteins in their native state, with secondary structure already formed, rather than studying the peptide bond in the unfolded state. Small peptides, which lack secondary structure, are believed to behave in similar ways to proteins in their unfolded state, and previous studies have found that standard molecular force fields often struggle to reproduce peptide QM results (21). The fact that the same peptide in different molecular force fields has different propensity to form helical or extended structures is a well known phenomenon (21, 25) and corrections to existing force fields to accurately reproduce helix propensity have been developed (25).

Galilean Nested Sampling

The Galilean nested sampling simulations in the Results Section used $\theta = 0.2$, which introduced a small amount of randomization at every Galilean step. This randomization is important in order to efficiently sample the system; Figure 10 (left) shows three nested sampling simulations with exactly the same parameters as those in Figure 3 except that $\theta = 0.01$ rather than 0.2. The same REMD data are shown for ease of comparison.

We also find that having a large number of short trajectories is beneficial, as shown by Figure 10 (right). In this figure, the same parameters were used as in Figure 3, except instead of 16000 trajectories of 2700 steps each, 160 trajectories of 270000 were used at each iteration.

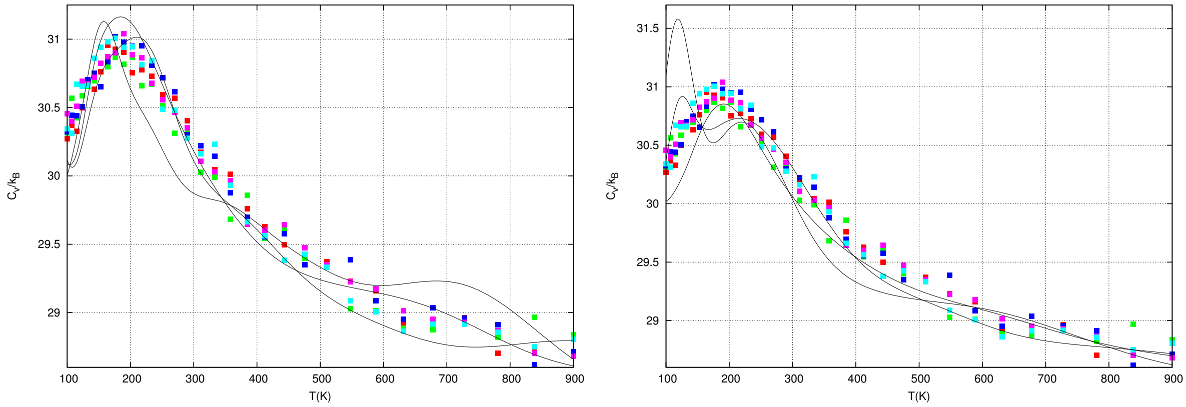


Figure 10: Nested sampling simulations of alanine dipeptide *in vacuo* with the same number of force evaluations as those in Figure 3. The same parameters were used except in left: $\theta = 0.01$ rather than 0.2 and in right: 160 trajectories of 270000 were used each iteration instead of 16000 trajectories of 2700 steps each. The REMD data from Figure 3 are shown for ease of comparison.

In previous studies, accurate heat capacities of alanine dipeptide have not been calculated, and because the curve is *almost* constant, varying by only $\sim 2k_B$ over the 800K temperature range, we believe that a large number of force evaluations are required in order to clearly resolve the curve. Figure 11 shows the heat capacity estimates using an order of magnitude fewer force evaluations[6]

[6] Note, though, that we did not reduce the initial equilibration period, as we did not want an unequilibrated initial set to affect the comparison.

for nested sampling, as compared to Figure 3. Figure 11 (left) reduces the number of trajectories of each nested sampling iteration by a factor of ten, and this clearly reduces the quality of the curves generated. Figure 11 (right) instead reduces the length of each trajectory by a factor of ten. Although the general shape of the heat capacity can still be resolved, individual C_v curves are of a lower quality than those of Figure 3. Once again, the previous REMD data is shown for ease of comparison between figures. It is important to note that $\approx 10^9$ force evaluations is still a large number of force evaluations for a system with only 60 internal degrees of freedom, of which very few (notably the dihedrals ϕ and ψ) are not highly constrained.

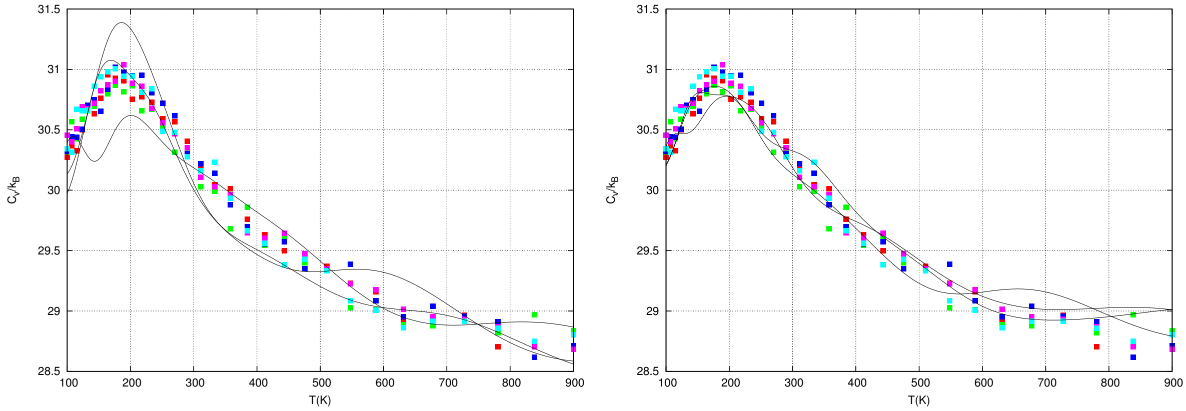


Figure 11: Heat capacity estimates from nested sampling simulations using an order of magnitude fewer force evaluations, as compared to Figure 3. Left: the number of trajectories of each nested sampling iteration by a factor of ten. Right: the length of each trajectory is reduced by a factor of ten.

In the future, we expect to test Galilean nested sampling with larger proteins. If a protein has a well-defined and known tertiary structure, i.e. a single dominant free energy minimum, then by starting all replicas of an REMD simulation from this minimum, the amount of equilibration is significantly reduced, as the protein does not need to be folded before investigating its thermodynamics. For example Yeh *et al.* (63) calculate the heat capacity of an SH3 domain in different implicit solvents, starting trajectories from the crystal structure of the protein. Unfortunately, the top-down nature of nested sampling does not easily allow simulations to be started from the crystal structure.

However, there has been a lot of recent interest in intrinsically disordered proteins (IDPs), that is proteins which do not have a well-defined fixed structure, which may, for example, only take well-defined structure upon binding. This interest is because it is now understood that they are significantly more common and important than first thought and perform a variety of biological functions, often related to human disease (64, 65). For these proteins, in equilibrium, there is a distribution over a set of possible macrostates, as is the case for alanine dipeptide. In this case, there is not a single obvious starting conformation for REMD replicas, and hence we believe nested sampling, with its top-down approach, will be particularly beneficial for the study of the thermodynamics of IDPs.

Behaviour at First-Order Phase Transitions

An alternative nested sampling algorithm has been introduced by Nielsen (58). In this work, canonical MC Markov chains, at a chosen (heat bath) temperature T_N, are generated. The samples output are then used to calculate the next nested sampled energy level using importance sampling; the samples are weighted by the inverse Boltzmann factor $\exp(+E(\Omega)/k_B T_N)$ in order to ‘simulate’ a uniformly distributed active set. Nielsen suggests that this algorithm could also be used with an MD framework, as canonical MD trajectories could be used in order to generate the required samples.

Due to the exponential Boltzmann factor, the standard deviation of the potential energies of the Boltzmann distribution is much larger than the standard deviation of uniformly distributed samples. Therefore, in order to improve the number of samples given non-negligible importance sampling weights, Nielsen adds an extra harmonic umbrella biasing term to narrow the standard deviation of the potential energy of samples.[7] The temperature T_N is controlled throughout the simulation in order to ensure that the energy of samples is found around the current nested sampling energy level.

In this initial exploratory study we focused on small peptides, relatively small systems which do not undergo a phase transition. However, nested sampling is expected to significantly outperform REMD, its extensions such as adaptive temperature REMD (27, 29), and other thermal sampling algorithms, when applied to high dimensional systems which undergo a first order phase transition. In this section we consider the behaviour of Galilean nested sampling, canonical nested sampling and REMD when they encounter a first order phase transition.

Figure 12 shows both the internal energy as a function of temperature and the evolution of the potential energy probability distribution as a function of temperature for a finite system undergoing a (quasi) first order phase transition at temperature T_c. In thermal equilibrium at T_c, the (relatively) small number of conformations of phase X with low potential energy ($\approx U_X$) exactly balance the large number of high energy ($\approx U_Y$) conformations of phase Y, and so overall the probability of the system being found in phase X is exactly a half.

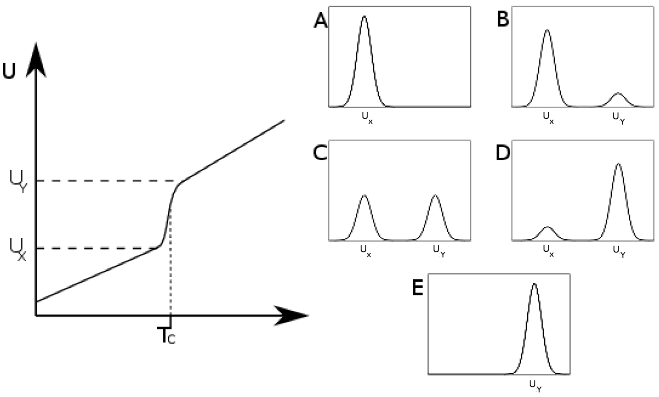


Figure 12: Left: The internal energy (U) as a function of temperature for a system exhibiting a first order transition at temperature T_c. Right: The potential energy probability distribution for this system at five different temperatures A) $T \ll T_c$; B) $T < T_c$; C) $T = T_c$; D) $T > T_c$; E) $T \gg T_c$.

Even with an REMD replica at T_c, it is difficult for REMD to equilibrate between the high energy, large phase space volume phase Y and the low energy, small phase space volume phase X. The larger the discrepancy of phase space volume between the phases, the harder the equilibration becomes.

It is important, however, to note that for finite systems it is possible to have a sigmoidal internal energy curve without a first order phase transition (66), i.e. the probability distributions of potential energy remain unimodal. In this case adaptive temperature REMD could be expected to perform well; a large number of replicas could be placed at the sigmoidal 'bend' and there would be no equilibration problem for individual replicas. In contrast, Galilean nested sampling, and similarly standard MC nested sampling, has no need to equilibrate between phases; at first order transitions, the algorithm steadily marches down in energy from U_Y to U_X, unaware that there was no temperature for which conformations with potential energy between U_X and U_Y were accessible.

Finally, we consider the canonical nested sampling of Nielsen. Due to the exponential weighting

[7] This bias is, of course, removed when calculating the next nested sampling energy level.

of samples, only those conformations with energy very close to the current energy level have a non-negligible weight, and therefore it is not necessary for samples from one iteration to equilibrate between different phases. However, a heat-bath temperature must be chosen for each iteration, even those when the energy level falls between U_X and U_Y. By applying umbrella biasing potentials, it may be possible to carefully control the temperature in order to retain the performance of nested sampling. However, this is likely to be very challenging as, even with a biasing potential, the temperature may not change monotonically throughout a simulation. Furthermore, if strong biasing potentials are applied, smaller integration steps are likely to be required, further reducing the performance of the algorithm. Finally, as few samples have non-negligible weights, a lot of computational expense is wasted exploring parts of the PES which are not currently relevant for the algorithm.

Comparing the efficiency of Galilean nested sampling, REMD and canonical nested sampling for sampling biophysical systems, such as proteins (54–56), and physical systems, such as water boxes, that are known to undergo first order phase transitions is a current focus of our work.

Conclusion and Future Work

In this study we have implemented Galilean nested sampling for use with the widely used Amber MD package. We have demonstrated the algorithm by sampling alanine dipeptide both *in vacuo* and using a generalized Born implicit solvent model. We have calculated heat capacity curves, and, by comparing our results with those generated by REMD, we have shown that it is possible to achieve good agreement between different sampling algorithms when estimating peptide heat capacity curves.

We have also calculated the Ramachandran free energy surface for the latest Amber force field, both *in vacuo* and with a generalized Born implicit solvent, and compared the results to previous theoretical and experimental work, notably the recent infra-red and Raman spectra data of Grdadolnik *et al.* (6).

We have discussed the performance of peptide force fields and considered in detail the behaviour of Galilean nested sampling for this system. We have also considered the theoretical behaviour of the algorithm, REMD and canonical nested sampling for systems which undergo a first order phase transition.

In this work we sampled Galilean velocities $\mathbf{v} = \mathbf{S}\mathbf{r}$ where $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{S} = \sqrt{k_B T}\mathbf{I}$ with the identity matrix $\mathbf{I}$. In the original description of the algorithm, Skilling suggests that certain choices of 'semimetric' $\mathbf{S}$ could be used to improve Galilean exploration (57). The reflection formula is then adapted to preserve detailed balance.

Specifically, Skilling suggests the semimetric $\mathbf{S} \approx (-\nabla \mathbf{F})^{-1/2}$, where $\mathbf{F}$ are the forces, at a preferred configuration (57). This semimetric takes into account the curvature of the space when choosing velocities. We believe this improvement would be essential for using Galilean nested sampling with larger molecular systems. This is because in molecular systems, certain degrees of freedom, such as the stretching of covalent bonds, are very highly constrained, whereas others, such as the dihedral angles ϕ and ψ, are not very constrained at all. It is clear that the magnitude of velocities in the highly constrained directions should be smaller than those in other directions in order to maximize efficiency. Preliminary results using the isotropic algorithm (i.e. $\mathbf{S} \propto \mathbf{I}$) for the penta-peptide Met-enkephalin suggest an appropriate semimetric would essential when using Galilean nested sampling with larger biophysical systems.

We conclude that Galilean nested sampling, with an appropriate semimetric, is a promising conformational sampling algorithm for biophysical atomistic systems, and we look forward to investigating its performance compared to other general-purpose sampling algorithms (i.e. those where no prior knowledge of the PES is required) such as REMD, accelerated MD and multicanonical MD, when sampling larger peptides and proteins.

References

[1] Ramachandran, G., C. T. Ramakrishnan, and V. Sasisekharan, 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7:95–99.

[2] Madison, V., and K. D. Kopple, 1980. Solvent-dependent conformational distributions of some dipeptides. J. Am. Chem. Soc. 102:4855–4863.

[3] Schweitzer-Stenner, R., F. Eker, Q. Huang, and K. Griebenow, 2001. Dihedral angles of trialanine in D2O determined by combining FTIR and polarized visible Raman spectroscopy. J. Am. Chem. Soc. 123:9628–9633.

[4] Pohl, G., A. Perczel, E. Vass, G. Magyarfalvi, and G. Tarczay, 2007. A matrix isolation study on Ac-Gly–NHMe and Ac-l-Ala–NHMe, the simplest chiral and achiral building blocks of peptides and proteins. Phys. Chem. Chem. Phys. 9:4698–4708.

[5] Schweitzer-Stenner, R., 2009. Distribution of conformations sampled by the central amino acid residue in tripeptides inferred from amide I band profiles and NMR scalar coupling constants. J. Phys. Chem. B 113:2922–2932.

[6] Grdadolnik, J., V. Mohacek-Grosev, R. L. Baldwin, and F. Avbelj, 2011. Populations of the three major backbone conformations in 19 amino acid dipeptides. Proc. Natl. Acad. Sci. USA 108:1794–1798.

[7] Jorgensen, W. L., and J. Gao, 1988. Cis-trans energy difference for the peptide bond in the gas phase and in aqueous solution. J. Am. Chem. Soc. 110:4212–4216.

[8] Head-Gordon, T., M. Head-Gordon, M. J. Frisch, C. Brooks, and J. Pople, 1989. A theoretical study of alanine dipeptide and analogs. Inter. J. Quantum Chem. 36:311–322.

[9] Tobias, D. J., and C. L. Brooks III, 1992. Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results. J. Phys. Chem. 96:3864–3870.

[10] Montgomery Pettitt, B., and M. Karplus, 1985. The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach. Chem. Phys. Lett. 121:194–201.

[11] Miao, Y., W. Sinko, L. Pierce, D. Bucher, R. C. Walker, and J. A. McCammon, 2014. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. J. Chem. Theory Comput. 10:2677–2689.

[12] Becker, O. M., and M. Karplus, 1997. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. 106:1495–1517.

[13] Wang, J., R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, 2004. Development and testing of a general amber force field. J. Comput. Chem. 25:1157–1174.

[14] Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., 2009. CHARMM: the biomolecular simulation program. J. Comput. Chem. 30:1545–1614.

[15] Antonczak, S., G. Monard, M. F. Ruiz-López, and J.-L. Rivail, 1998. Modeling of peptide hydrolysis by thermolysin. A semiempirical and QM/MM study. J. Am. Chem. Soc. 120:8825–8833.

[16] Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim, and M. Karplus, 2001. A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method. J. Phys. Chem. B 105:569–585.

[17] Oie, T., G. H. Loew, S. K. Burt, J. S. Binkley, and R. D. MacElroy, 1982. Quantum chemical studies of a model for peptide bond formation: formation of formamide and water from ammonia and formic acid. J. Am. Chem. Soc. 104:6169–6174.

[18] Evans, D. A., and D. J. Wales, 2004. Folding of the GB1 hairpin peptide from discrete path sampling. J. Chem. Phys. 121:1080–1090.

[19] Yeh, I.-C., and A. Wallqvist, 2009. Structure and dynamics of end-to-end loop formation of the penta-peptide Cys-Ala-Gly-Gln-Trp in implicit solvents. J. Phys. Chem. B 113:12382–12390.

[20] Hughes, J., T. Smith, H. Kosterlitz, L. Fothergill, B. Morgan, and H. Morris, 1975. Identification of two related pentapeptides from the brain with potent opiate agonist activity. Nature 258:577–579.

[21] Hu, H., M. Elstner, and J. Hermans, 2003. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine “dipeptides” (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. Proteins: Struct. Funct. Genet. 50:451–463.

[22] Nguyen, H. D., and C. K. Hall, 2004. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc. Natl. Acad. Sci. USA 101:16180–16185.

[23] Banks, J. L., G. A. Kaminski, R. Zhou, D. T. Mainz, B. J. Berne, and R. A. Friesner, 1999. Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model. J. Chem. Phys. 110:741–754.

[24] Winther, O., and A. Krogh, 2004. Teaching computers to fold proteins. Phys. Rev. E 70:030903.

[25] Best, R. B., and G. Hummer, 2009. Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides. J. Phys. Chem. B 113:9004–9015.

[26] Sugita, Y., and Y. Okamoto, 1999. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 314:141–151.

[27] Katzgraber, H. G., S. Trebst, D. A. Huse, and M. Troyer, 2006. Feedback-optimized parallel tempering Monte Carlo. J. Stat. Mech. Theor. Exp. 2006:P03018.

[28] Simmerling, C., B. Strockbine, and A. E. Roitberg, 2002. All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. 124:11258–11259.

[29] Lee, M. S., and M. A. Olson, 2011. Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. J. Chem. Phys. 134:244111.

[30] Torrie, G. M., and J. P. Valleau, 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J. Comput. Phys. 23:187–199.

[31] Straatsma, T., and H. Berendsen, 1988. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. J. Chem. Phys. 89:5876–5886.

[32] Kong, X., and C. L. Brooks III, 1996. λ-dynamics: A new approach to free energy calculations. J. Chem. Phys. 105:2414–2423.

[33] Tobias, D. J., S. F. Sneddon, and C. L. Brooks III, 1990. Reverse turns in blocked dipeptides are intrinsically unstable in water. J Mol. Biol. 216:783–796.

[34] Hamelberg, D., J. Mongan, and J. A. McCammon, 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. 120:11919–11929.

[35] Berg, B. A., and T. Neuhaus, 1992. Multicanonical ensemble: A new approach to simulate first-order phase transitions. Phys. Rev. Lett. 68:9–12.

[36] Hansmann, U. H., Y. Okamoto, and F. Eisenmenger, 1996. Molecular dynamics, Langevin and hydrid Monte Carlo simulations in a multicanonical ensemble. Chem. Phys. Lett. 259:321–330.

[37] Nakajima, N., J. Higo, A. Kidera, and H. Nakamura, 2000. Free energy landscapes of peptides by enhanced conformational sampling. J. Mol. Biol. 296:197–216.

[38] Higo, J., N. Ito, M. Kuroda, S. Ono, N. Nakajima, and H. Nakamura, 2001. Energy landscape of a peptide consisting of α-helix, 310-helix, β-turn, β-hairpin, and other disordered conformations. Prot. Sci. 10:1160–1171.

[39] Higo, J., J. Ikebe, N. Kamiya, and H. Nakamura, 2012. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. Biophys. Rev. 4:27–44.

[40] Wang, F., and D. P. Landau, 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys. Rev. Lett. 86:2050–2053.

[41] Kou, S., Q. Zhou, W. H. Wong, et al., 2006. Equi-energy sampler with applications in statistical inference and statistical mechanics. Ann. Stat. 34:1581–1619.

[42] Bonomi, M., and M. Parrinello, 2010. Enhanced sampling in the well-tempered ensemble. Phys. Rev. Lett. 104:190601.

[43] Skilling, J., 2004. Nested sampling. AIP Conf. Proc. 735:395–405.

[44] Do, H., J. D. Hirst, and R. J. Wheatley, 2011. Rapid calculation of partition functions and free energies of fluids. J. Chem. Phys. 135:174105.

[45] Do, H., J. D. Hirst, and R. J. Wheatley, 2012. Calculation of partition functions and free energies of a binary mixture using the energy partitioning method: application to carbon dioxide and methane. J. Phys. Chem. B 116:4535–4542.

[46] Skilling, J., 2006. Nested Sampling for General Bayesian Computation. J. Bayesian Anal. 1:833–859.

[47] Mukherjee, P., D. Parkinson, and A. R. Liddle, 2006. A nested sampling algorithm for cosmological model selection. Astrophys. J. Lett. 638:L51.

[48] Doğruel, M., T. A. Down, and T. J. Hubbard, 2008. NestedMICA as an ab initio protein motif discovery tool. BMC Bioinformatics 9:19–31.

[49] Pullen, N., and R. J. Morris, 2014. Bayesian Model Comparison and Parameter Inference in Systems Biology Using Nested Sampling. PLOS ONE 9:e88419.

[50] Elsheikh, A., M. Wheeler, and I. Hoteit, 2013. Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration. Water Resources Research 49:8383–8399.

[51] Pártay, L. B., A. P. Bartók, and G. Csányi, 2010. Efficient sampling of atomic configurational spaces. J. Phys. Chem. B 114:10502–10512.

[52] Pártay, L. B., A. P. Bartók, and G. Csányi, 2014. Nested sampling for materials: The case of hard spheres. Phys. Rev. E 89:022302.

[53] Burkoff, N. S., C. Várnai, S. A. Wells, and D. L. Wild, 2012. Exploring the energy landscapes of protein folding simulations with Bayesian computation. Biophys. J. 102:878–886.

[54] Shakhnovich, E. I., and A. V. Finkelstein, 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. Biopolymers 28:1667–1680.

[55] Privalov, P., and N. Khechinashvili, 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. J. Mol. Biol. 86:665–684.

[56] Nojima, H., A. Ikai, T. Oshima, and H. Noda, 1977. Reversible thermal unfolding of thermostable phosphoglycerate kinase. Thermostability associated with mean zero enthalpy change. J. Mol. Biol. 116:429–442.

[57] Skilling, J., 2012. Bayesian computation in big spaces-nested sampling and Galilean Monte Carlo. *In* Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Publishing, Melville, NY, USA, Volume 1443, 145–156.

[58] Nielsen, S. O., 2013. Nested sampling in the canonical ensemble: Direct calculation of the partition function from NVT trajectories. J. Chem. Phys. 139:124104.

[59] Raftery, A. E., M. A. Newton, J. M. Satagopan, and P. N. Krivitsky, 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Oxford University Press, Oxford, UK.

[60] Podtelezhnikov, A. A., and D. L. Wild, 2008. CRANKITE: A fast polypeptide backbone conformation sampler. Source Code Biol. Med. 3:1–7.

[61] Still, W. C., A. Tempczyk, R. C. Hawley, and T. Hendrickson, 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112:6127–6129.

[62] Wang, Z.-X., W. Zhang, C. Wu, H. Lei, P. Cieplak, and Y. Duan, 2006. Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. J. Comput. Chem. 27:781–790.

[63] Yeh, I.-C., M. S. Lee, and M. A. Olson, 2008. Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. J. Phys. Chem. B 112:15064–15073.

[64] Dunker, A. K., J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, 2001. Intrinsically disordered protein. J. Mol. Graphics Modell. 19:26–59.

[65] Eliezer, D., 2009. Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 19:23–30.

[66] Finkelstein, A. V., and O. Ptitsyn, 2002. Protein physics: a course of lectures, Academic Press, San Diego, CA, USA, Chapter 8.

Supporting Material

Starting the Algorithm at a Chosen Initial Energy Level

In this work we choose an initial energy level, E_0, rather than sampling the parts of conformational space with high energy. This improves the efficiency of the algorithm as fewer iterations are needed. Provided E_0 is chosen to be higher than accessible conformations over the range of temperatures of interest, choosing an initial energy level does not affect thermodynamic estimates at these temperatures.

As an example, although our estimate for $\log Z(\beta)$ is out by an additive constant,[1] its derivatives (w.r.t. β), the internal energy and heat capacity, are unaffected by this constant. Similarly, when calculating free energy differences using Equation (3), this constant cancels out.

It is extremely important that E_0 is high enough to allow easy equilibration. For example, if at E_0 the acceptable region was disconnected, then it would be very challenging to initially equilibrate an active set. Figure S1 shows the dihedral angles of the samples output by a nested sampling simulations of alanine dipeptide *in vacuo* which have potential energy between 44 and 50 kcal/mol (left) and between 86 and 100 kcal/mol (right). It is possible for samples to cross the potential energy barrier with the energy constrained to below 50 kcal/mol. However it is significantly easier if the energy is allowed to rise to 100 kcal/mol, and this is the value we chose for E_0.

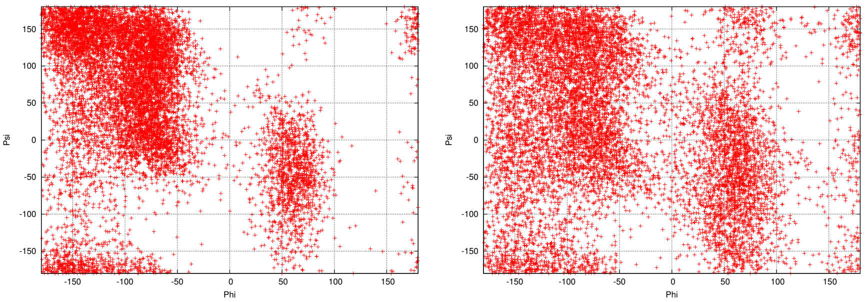


Figure S1: The dihedral angles of conformations output by nested sampling simulations which have energy between 44 and 50 kcal/mol (left) and between 86 and 100 kcal/mol. Both graphs contain approximately 10^4 data points.

With regards to canonical molecular dynamics, at 900K the potential energy distribution is approximately Gaussian, with mean ≈ 31 and standard deviation ≈ 10, which allows the trajectories to cross the barrier. At 600K the distribution has mean ≈ 13 and standard deviation ≈ 6, making it very difficult to cross the barrier. Therefore, it is likely that incorrect thermodynamics would have been calculated if the highest temperature replica was only 600K.

In order to generate an active set of conformations uniformly distributed over configuration space, subject to having energy $< E_0$, we first generate a set of independent conformations with uniformly distributed dihedral angles. We then perform short minimizations so their energies $< E_0$.

At this point, it is possible to use Galilean exploration to equilibrate the samples. In this work, however, for initialization, although we perform Galilean exploration, rather than reflecting at the boundary, we allow samples to cross the boundary and then perform standard *canonical* exploration until we re-enter the allowed region. This is not strictly necessary, but does ensure the samples are well equilibrated and, if desired, could allow a lower E_0 to be used than otherwise would be possible.

The total number of force evaluations used in the initial equilibration is not included in the comparison with REMD in the main text, yet is only of the order of a couple of nested sampling

[1] If E_0 is chosen so that we miss the first j nested sampling iterations and β is large enough so that $\sum_i^j \omega_i \exp(-E_i\beta)$ is negligible, using Equation (1), our estimate for Z is out by the multiplicative factor ω_j.

iterations.

Temperature Control

As the accessible region of phase space shrinks, it is necessary to reduce the magnitude of the velocities in order to keep the trajectories within the allowed region. At each iteration, the accessible volume (V) of phase space shrinks by a factor α. Therefore, for a system of dimension d, the length scale shrinks by a factor of $\alpha^{1/d}$ and, as $T \sim <\text{speed}>^2$, this implies we should multiply T by a factor of $\alpha^{2/d}$ each iteration in order to maintain a chosen mean free path.

Alanine dipeptide is a 22-atom molecule. Hence our (non-periodic) system has 60 internal degrees of freedom and so we multiply T by a factor of $0.5^{2/60}$ each iteration. The initial T is chosen empirically, tuned to generate the desired mean free path. The timestep τ is fixed at 0.2fs throughout.

The theoretical argument above does not take into account the changing shape of the phase space throughout the nested sampling simulation, so for more complicated potential energy surfaces it may be necessary to algorithmically adjust T in order to control the mean free path. However, for alanine dipeptide, although there is a slight drift, the simple control is good enough, as shown by Figure S2.

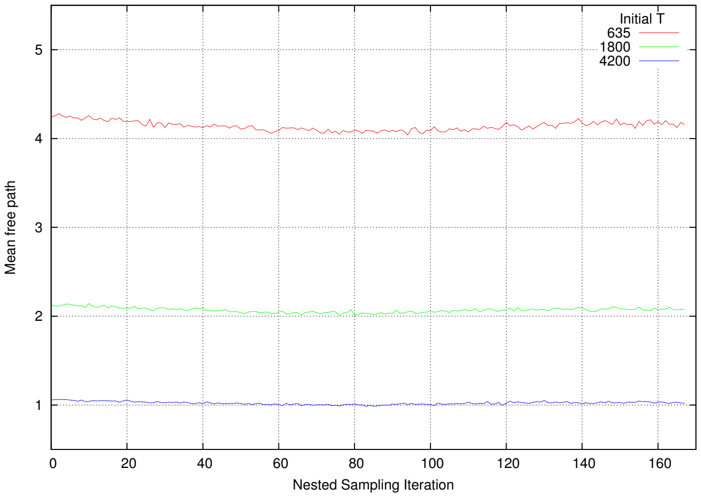


Figure S2: The mean free path throughout alanine dipeptide vacuum nested sampling simulations using three different initial T and reducing T by a factor of $0.5^{2/60}$ each iteration. The mean free path remains fairly stable over the whole simulation. The proportion of Galilean reflections accepted was $\approx 0.98, 0.97$ and 0.93 for mean free paths 4, 2 and 1 respectively.

Unsmoothed Free Energy Surface

A Gaussian smoothing function has been applied to the raw data when generating the free energy surfaces in the main text (see for example Figures 4). For comparison, Figure S3, below, shows the raw unsmoothed data for alanine dipeptide *in vacuo* at 300K for three of the independent simulations whose heat capacity curves are shown in Figure 3.

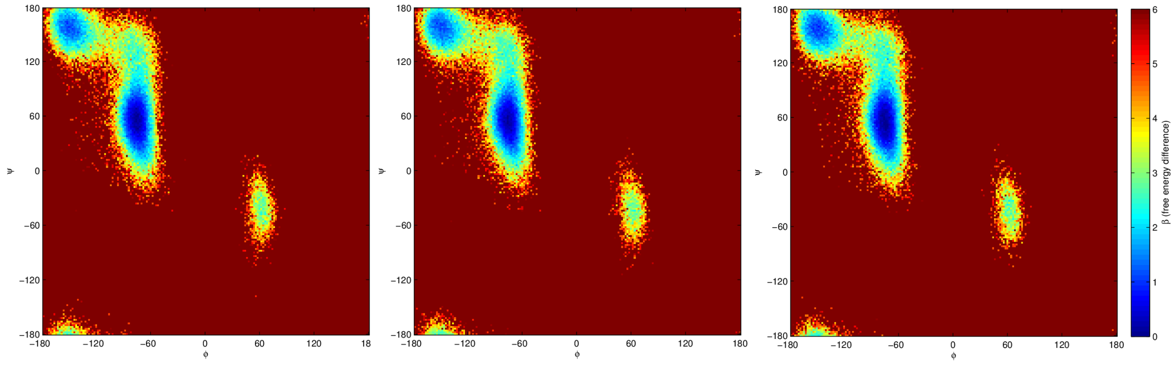


Figure S3: The raw (unsmoothed) free energy surfaces of alanine dipeptide *in vacuo* at 300K from three independent nested sampling simulations.

Chapter 6

Conclusions

6.1 Thesis Summary

6.1.1 Nested Sampling of Proteins and Peptides

Nested sampling is a Bayesian algorithm designed to be particularly efficient at sampling systems which undergo a first order phase transition. In Chapter 2 we parallelized the algorithm and, for the first time, used it to sample a biophysical system, a coarse-grained protein model, CRANKITE.

The potential energy landscapes of three small proteins are explored and energy landscape charts are generated, giving a large-scale visualization of the potential energy surface showing the protein folding funnel. We considered how the simulations behave when changing the NS algorithm parameters. We also compared the nested sampling algorithm to parallel tempering, using both methods to calculate the heat capacity of polyalanine.

For more complicated protein models, which have more degrees of freedom per residue, the MC move set must allow for, for example, angle bending and side chain rotations. These moves, especially at low temperatures, are often inefficient when compared to the more widely used molecular dynamics approach. This will be particularly noticeable in cases where explicit solvent molecules are included, which is often the case for biophysical systems.

When using explicit solvent molecules with MD, collisions between separate solvent molecules exchange energy and can enable energy barriers to be crossed, whereas with MC, a large number of moves cause molecules to overlap, which can cause a low acceptance rate and long decorrelation times. Therefore, in order for NS to gain popularity within the computational structural biology community, it is necessary to adapt the algorithm to work within an MD framework.

To that end, in Chapter 5, we adapted the nested sampling algorithm to be used within an MD framework by implementing Galilean exploration. We demonstrated the application of the algorithm by calculating heat capacity curves for an all-atom model of alanine dipeptide and compared the results to the standard replica exchange approach. We calculated the dihedral angle free energy surface of alanine dipeptide both *in vacuo* and implicit solvent and used the surface to compare the latest Amber force field to previous computational and experimental work.

Finally, we discussed the theoretical behaviour of Galilean nested sampling, REMD and an alternative nested sampling algorithm, which uses canonical trajectories, for systems which undergo a first order phase transition. After incorporating an appropriate semimetric, Galilean exploration should allow NS to be used with more realistic force fields where there is often no efficient MC move set.

6.1.2 Contrastive Divergence and Protein Force Field Parameter Optimization

In this work we have substantially improved CRANKITE, a coarse-grained protein model. In Chapter 2 we added side-chain γ-atoms to the model, together with an MC side-chain dihedral angle rotation move. We also improved the energy function by adding a hydrophobic energy term and tuning the functional forms of existing energy terms.

In Chapter 3 we focussed on optimizing the parameters of the CRANKITE force field. We used a maximum likelihood approach, optimizing the force field parameters such that the likelihood of a training set, consisting of experimentally-derived protein crystal structures, is maximized.

In order to avoid the expensive calculation of ensemble averages, we used a statistical machine-learning technique, contrastive divergence. In comparison to other maximum likelihood approaches, the efficiency of our algorithm allows a larger training set to be used and we have shown the optimized force field is transferable to a protein not included in the training set.

In Chapter 3 we placed particular emphasis on the van der Waals energy term. We optimized parameter values for both a cheap, hard cutoff function and a more expensive 12–6 LJ functional form, and we compared them to parameters taken from 'standard' molecular dynamics force fields: we compared the observed distributions of bond angles, atomic distances, backbone dihedral angles and hydrogen bonding patterns. We also calculated the heat capacities of polyalanine and observed the different turn types found when folding a β-hairpin. We demonstrated the importance of optimizing the parameters of the force field rather than taking values found in the literature.

In Chapter 3 we also discussed the contrastive divergence procedure as applied to force field parameter inference, its behaviour, the assumptions it relies upon and the effect of changing the quality of the training set.

6.1.3 β-Contact Prediction and Correlated Mutation Analysis

In Chapter 4 we developed a protein β-contact prediction algorithm whose predictions can be used as inputs to CRANKITE when the native protein structure is unknown. We developed an empirical Bayes β-sheet model which encodes the strong constraints and prior knowledge associated with β-contacts. We coupled the model to the direct information (DI), a powerful maximum entropy-based correlated mutation statistic.

Unlike the majority of correlated mutation analysis research, proteins with large high-quality multiple sequence alignments are not specifically chosen for analysis, but instead a standard dataset of 916 proteins used to benchmark β-contact prediction algorithms is used. We show that the DI statistic contains useful information even when smaller autogenerated MSAs are used, and that, according to our benchmarked results, the DI is as informative as inputting the *entire* MSA into a neural network or Markov random field when predicting β-contacts.

Finally, tying this work in with the rest of the thesis, we show that the β-contact predictions can be used within a tertiary structure prediction pipeline by using them as inputs to CRANKITE, enabling it to successfully determine the folds of two previous CASP targets.

6.2 Discussion and Future Work

In this section possible future projects, extending the work described in this thesis, are discussed. Further applications of contrastive divergence to force field parameter optimization are considered and the future of coarse-grained protein models more generally is discussed. Possible improvements to the β-contact prediction algorithm, and its use within tertiary prediction schemes, are explored. The use of correlated mutation analysis for predicting protein-protein interactions is also examined. However, first, improvements to, and further applications of, the nested sampling algorithm are considered.

6.2.1 Improving the Nested Sampling Algorithm

Multimodal Potential Energy Landscapes

As described in Chapter 2, the copying procedure used by nested sampling to generate starting conformations for the MC chains has a distinct advantage for multimodal systems such as the example shown in Figure 6.1 (left). Basin B splits into basins C and D at energy level E, a lower energy than the 'bottom' of basin A. Provided one sample 'finds' basin B, the copying procedure will ensure all K samples in the active set are found in basin B at energy E, thus maximizing the chances of the algorithm finding basin D.

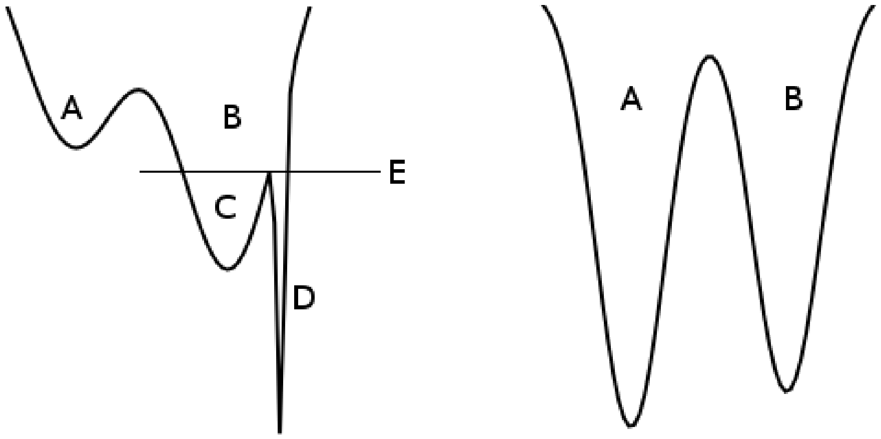


Figure 6.1: Left: A multimodal energy surface for which the copying procedure of nested sampling is advantageous. Provided one sample 'finds' basin B, the copying procedure will ensure all K samples in the active set are found in basin B at energy E, thus maximizing the chances of the algorithm finding basin D. Right: An energy suface for which the copying procedure is not particularly useful. See the text for further details.

However, if a system has a potential energy surface which resembles Figure 6.1 (right), such as protein G using CRANKITE (Figure 3 Chapter 2), the copying procedure is not especially advantageous. Once the nested sampling energy level is lower than the energy of separation for the basins, the algorithm is sampling a disconnected space and it is relying on the copying procedure to equilibrate between the modes. The copying procedure is a random diffusive process and if the number of nested sampling iterations though the disconnected space required to reach the posterior mass is large, then it is likely that no samples will be left in one of the basins and the evidence value will be lost.

To overcome this problem, an enormous active set must be used, which means a tiny fraction of available phase space is removed in each iteration, leading to a substantial loss of efficiency. It is also worth mentioning that, although CRANKITE is a coarse-grained model, similar behaviour, i.e. a small number of deep basins separated by high energy barriers, is also found when exploring the PES of alanine dipeptide using the all-atom Amber force field (see Chapter 5, Figures 4, 5 and S1).

Brewer *et al.* have described a possible solution to this problem with their diffusive nested sampling algorithm (225). In this procedure, instead of a hard energy cutoff, the MC chain is allowed to diffuse up to higher energy levels. However, for high energy barriers, the problem is likely to remain.

An alternative solution is to identify the modes during the nested sampling simulation, for example by clustering the active set of samples and keeping track of whether samples change clusters after running the MC chains. Once modes have been identified, separate nested sampling simulations can be run independently, one for each mode, and then the individual results can be combined. A 'clustered' nested sampling algorithm has been developed for astrophysics parameter inference (226), but has not been implemented for atomistic systems.

The algorithm could be improved further by allowing the number of samples in the active set, and the length of the MC chains to vary throughout the simulation. For example, at high energies, where multimodality is not an issue and large MC steps can be accepted, shorter MC chains could be used. Similarly, at the end of a simulation, we essentially have K copies of the same conformation, and we slowly keep lowering the energy of all K samples. At this point, a much smaller active set could be used.

Galilean Nested Sampling

In Chapter 5 we implemented Galilean nested sampling for atomistic systems and demonstrated its validity by sampling alanine dipeptide. In that work we sample Galilean velocities $\mathbf{v} = \mathbf{S}\mathbf{r}$ where $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{S} = \sqrt{k_B T}\mathbf{I}$ with the identity matrix $\mathbf{I}$. In the original description of the algorithm, Skilling suggests that certain choices of 'semimetric' $\mathbf{S}$ could be used to improve Galilean exploration (224). The reflection formula is then adapted to ensure detailed balance is maintained.

Skilling suggests the semimetric $\mathbf{S} \approx (-\nabla\mathbf{F})^{-1/2}$, where $\mathbf{F}$ are the forces $(-\nabla E)$ at a preferred configuration (224). This semimetric takes into account the curvature of the space when choosing velocities, and it is hoped that the extra expense in its calculation is less significant than the benefits it brings, especially as an approximation to the semimetric can be used and this approximation need only be accurate enough to be useful.

As described in the conclusion of Chapter 5, we believe this improvement is essential for using Galilean nested sampling with larger molecular systems. This is because in molecular systems, certain degrees of freedom, such as the stretching of covalent bonds, are very highly constrained, whereas others, such as the dihedral angles ϕ and ψ, are not very constrained at all. It is clear that the magnitude of velocities in the highly constrained directions should be smaller than those in other directions and this is ensured by using the above choice of semimetric.

We do not envisage any serious difficulties in extending our Amber implementation of Galilean nested sampling in order to incorporate this semimetric; most of the routines that are required in order to calculate and diagonalize the Hessian, are available within the Amber source code.[1] An important consideration, however, is that detailed balance requires a constant $\mathbf{S}$ throughout a trajectory,[2] and therefore a preferred location must be chosen for the curvature calculation. When using internal co-ordinates, such as bond lengths and valence angles, certain directions (i.e. dimensions) are highly constrained throughout the entire PES. In this case the choice of preferred location is unlikely to be very important. However, in Cartesian co-ordinates, as used by Amber, in different parts of phase space, different direc-

[1] Though somewhat obfuscated.

[2] Though it can be allowed to vary occasionally (224).

tions relative to the reference axes are highly constrained. This may cause problems and may therefore increase the amount of parameter testing and validation time that is required. Alternatively, it may be preferable to work with internal co-ordinates, using an internal co-ordinate MD approach, as has been developed by Abagyan *et al.* (227). For further details, we refer the reader to the Appendix, which contains preliminary results demonstrating the effect of choosing a sensible semimetric.

It is interesting to note that at no stage are the masses of the particles used. As the PES is independent of particle mass, this is not especially surprising. However, we note that the semimetric $\mathbf{S} = \mathrm{diag}(\sqrt{k_BT/m_{1,1}}, \ldots, \sqrt{k_BT/m_{3,N}})$ where $m_{i,j}$ is the mass of particle j (and i runs from 1 to 3) for Galilean exploration would imply the velocities are picked according to a Maxwell–Boltzmann distribution, as is the case for canonical exploration. However, we do not believe this semimetric will be as useful as one which incorporates the curvature of the PES.

After implementing an appropriate semimetric, and possibly other algorithmic improvements described above, the next step is to test the algorithm on larger biophysical systems, first trying commonly tested small peptides such as Met-enkephalin (19) and Cys-Ala-Gly-Gln-Trp (228); then moving on to mini-proteins, such as the Trp-cage (119), and finally to small proteins such as the SH3 domain (135).

Due to the nature of larger peptides and proteins, the PESs of these systems are likely to have a number of deep basins, each basin containing a set of conformations with broadly similar backbone dihedral angles, and therefore the clustered nested sampling approach described above is likely to be effective. Care must also be taken to ensure the initial set of conformations are sufficiently equilibrated and that an initial energy level, if used, is sufficiently high to allow adequate sampling. See the Supporting Material in Chapter 5 for further details.

As described in Chapter 5, there is no theoretical reason against using Galilean nested sampling with explicit solvent molecules, and it could be interesting to use Galilean nested sampling to compare the free energy surface of alanine dipeptide in implicit and explicit solvent.

Comparing the performance of both standard MC nested sampling and Galilean nested sampling for sampling explicitly solvated proteins and peptides with other state-of-the art general-purpose sampling algorithms is a particularly exciting prospect. By general-purpose, we mean algorithms which do not rely on prior knowledge of the PES, such as REMD, accelerated MD (229) and multicanonical sampling (23, 230), rather than algorithms which require some additional knowledge such as umbrella sampling (33), where reaction co-ordinates are chosen *a priori*.

A comparison of nested sampling and multicanonical sampling algorithms is of particular interest as the multicanonical approach was also designed to efficiently sample systems which undergo a first order phase transition (23).

Choice of force field

As described above, many popular MD force fields allow bond lengths and valence angles to vary around their equilibrium value. This can be important, as this variation is found in experimentally determined structures (159). Even the peptide bond is often far from planar (231). However, allowing bond lengths and angle values to vary substantially increases the cost of energy evaluations. Therefore alternative, fixed bond and valence angle force fields have been developed.

One such example is the 'ultra-fast' (unbiased) implicit solvent PROFASI force field (232), which has been used with parallel tempering in order to study the folding of Top7, a designed protein (233). It is interesting to note that the authors could not accurately determine the peak of the heat capacity curve for this system. Also, when starting from the unfolded state, although structures with RMSD ≈ 3.5Å from the native structure were found, so were a large number of very stable non-native states. This may imply that the sampling was inadequate, and possibly that the force field is not able to determine the native structure to a high degree of accuracy. This suggest that it may be worth using nested sampling to fully explore the PES for proteins when using force fields such as PROFASI.

Assessing the accuracy of currently used force fields for protein folding studies is an active area of research, and the latest perspectives can be found in (129, 234). The results of Chapters 2 and 5 indicate how nested sampling may provide insights into the properties of particular protein models and force fields.

Further applications

In this work we do not find the orders of magnitude efficiency gains using nested sampling to accurately calculate thermodynamic averages which have been found previously (43). We believe this is because the systems we have studied either do not undergo first order phase transitions, or thermal methods manage to equilibrate between the phases (e.g. in Figure 1.2 of the thesis introduction $U_x \sim U_y$ and the regions of conformational space associated with each phase have similar volumes).[3]

Nested sampling is expected to perform well in systems with first order phase transitions (39), and therefore we would expect to see large efficiency gains when using nested sampling to calculate heat capacities of larger proteins because there is theoretical (74) and experimental evidence (77) (for example, calorimetric studies of lysozyme heat denaturation (76)) which suggests proteins undergo a first order phase transition upon folding.

Currently it would be very computationally expensive to perform such calculations, starting from the unfolded state, because of the large number of energy evaluations required to initially fold the protein, and also the fact that energy evaluations themselves become increasingly expensive when larger systems are studied. Furthermore, the top-down nature of standard nested sampling makes it unsuitable to start the system in the folded state. However, future increases in computational power, including GPU enabled MD packages (235, 236) for fast energy evaluations and special-purpose hardware (237), should enable nested sampling to be used for larger systems in the coming years.

Recently there has been a lot of interest in the study of intrinsically disordered proteins, that is, proteins without well-defined tertiary structure (at near-physiological conditions), because it is now understood that they are significantly more common and important than first thought (63, 64). When a protein has a well-defined (and known) tertiary structure, REMD can be used to investigate the thermodynamics of the protein starting each replica from the native structure, see for example (134). In this case the system is already closer to equilibrium than starting from an unfolded chain, and hence the computational expense should be dramatically reduced.

For intrinsically disordered proteins this is not possible, as the set of probable conformations under the low temperature Boltzmann distribution do not all share a well-defined tertiary structure, but are an equilibrated distribution of allowed macrostates, similar to the distribution of alanine dipeptide shown in Chapter 5 (Figure 9). For these proteins, it is not clear what the initial configurations should be for REMD *a priori*, and hence nested sampling should be particularly useful in this case.

The study of intrinsically disordered proteins is a field for which many traditional structural biology tools are not ideal and new tools are required. A recent review considers the latest models and MD algorithms for studying intrinsically disordered proteins which take a specific tertiary structure upon binding (238). Of particular interest is the work of Ozenne and colleagues (239), who have developed a

[3] In Chapters 2 and 3 we run nested sampling simulations of protein G using CRANKITE. However, we do not attempt to calculate the heat capacity curve for this system. It is important to note that the long-range harmonic potential which pulls β-strands together used in CRANKITE may affect the nature of the folding 'transition'.

computational approach to constructing equilibrium ensembles of intrinsically disordered proteins which are consistent with chemical shifts or residual dipolar couplings NMR experimental data. Their approach may allow more accurate force fields for intrinsically disordered proteins to be developed in the future.

Changing our view from 'the native structure of a protein' to 'the equilibrium distribution of the system', with regards to experiments and also when developing computational models and sampling algorithms, is an important step towards understanding the behaviour of intrinsically disordered biophysical systems.

Finally, in Chapter 5 we showed the calculation of free energy differences between different macrostates within the same system (alanine dipeptide). In principle, nested sampling can also be used to calculate free energy differences between different systems, for example, between a wild type protein and a mutant or between two different ligands binding to a protein. Separate nested sampling simulations could be run for each system and their partition functions compared.

Specifically, the *absolute* free energy of a system is given by $-\beta^{-1} \log Z$ and the free energy difference between two systems, A and B, is therefore $-\beta^{-1}(\log Z_A - \log Z_B)$. However, as the system size increases, it becomes increasingly difficult to accurately estimate the partition function. As only the free energy difference is required, direct estimation of the partition function is not attempted and a different approach is usually taken.

Many advanced 'alchemical' free energy algorithms have been developed in order to calculate absolute binding affinities (for example, the free energy difference between a bound and unbound protein and ligand) and relative binding affinities (for example, the free energy difference between a protein bound to ligand A or ligand B); see (240) and references therein for examples.

It is possible that nested sampling will not be an efficient algorithm for these calculations as it is appropriate to start simulations with already folded proteins, rather than exploring the *whole* PES. However, alchemical free energy algorithms often perform poorly when there are long-time scale effects or multiple modes of binding, because in these cases the sampling algorithms often to struggle to effectively sample the space (240). Even if nested sampling itself is unsuitable, this suggests that Galilean sampling could be incorporated into existing algorithms, instead of using standard MD trajectories, perhaps yielding substantial efficiency gains.

6.2.2 Contrastive Divergence and Coarse-Grained Protein Models

There are several possible improvements to the CRANKITE force field which should improve the accuracy and utility of the model, whilst allowing it to remain a lightweight, efficient, backbone conformational sampler. The addition of a (non-hydrogen bond) electrostatic energy term would allow the force field to capture protein salt bridges, and the incorporation of main chain–side chain hydrogen bonds may improve the stability of tertiary structures.

Implementing normal mode projections (such as those used to perform the rigidity analysis in Chapter 2) as possible 'moves' would allow large-scale conformational changes which could improve sampling efficiency, especially towards the end of nested sampling simulations. However, care must be taken in order to preserve detailed balance.

For protein structure prediction, it would be interesting to implement a disulphide bridge energy term which would mimic the formation (and persistence) of disulphide bridges. Similarly to the β-contact case, disulphide bridges could then be predicted (see, for example, (186)) and entered as inputs to CRANKITE.

For CRANKITE, the β-contact energy term is the main long-range 'force' which pulls the secondary structure elements together, though once the β-strands are close, the hydrogen bond energy term is then important (see Figure S4 in Chapter 2). This long-range term is essential when starting from an extended chain as it enables the protein to actually fold using a realistic amount of computer power.

However, for α-helical proteins, or proteins with few β-contacts, although the secondary structure elements form, the protein does not fold up into a globule. One possible solution is to try and predict more protein contacts and use them as inputs for CRANKITE.

An alternative solution is to use an energy term to pull the secondary structure elements together *whilst still allowing them to form.* The footnote of the supplementary data of Chapter 4 describes such an example: the energy term constrains the radius of gyration of the centre of masses of the distinct helices and strands (whose residues are inputs to CRANKITE). Further work will be needed to fully test this new energy term.

The reference ratio method of Hamelryck *et al.* (241) is a statistically rigorous method which aims to take force fields which accurately model 'local' protein properties, such as steric attractions and secondary structure elements, and use them to sample proteins with user-specified distributions of 'global' properties, such as radius of gyration and hydrogen bond networks. Implementing this method could be an alternative approach to improve the quality of tertiary structures generated by CRANKITE.

Contrastive Divergence and the Maximum Likelihood Approach

Contrastive divergence, and the maximum likelihood approach more generally, can be used whenever parameters are to be inferred from a training set. The approach is agnostic to model-specific details. A particularly interesting example of the approach is the 2013 work by Bottaro *et al.*. They use a maximum likelihood (also known as relative entropy) method to optimize an implicit solvent model using a training set of configurations derived from an all-atom explicit solvent model. The aim of the work is developing a solvent model which retains the accuracy of an explicit solvent model, yet computationally is significantly cheaper (242).

The maximum likelihood method has already been used in a framework for protein design (243). However, in the future it would be interesting to use the contrastive divergence approximations to optimize force fields specifically developed for aiding protein design or protein-protein docking prediction.

Additionally, it may also be possible to develop a maximum likelihood framework for picking native structures from a set of decoys (protein-like configurations). For example, a binary classifier could be developed to separate native structures from decoys, and the parameters of the model could be optimized within the maximum likelihood framework. This would be of enormous use to the template-based protein structure prediction community as often these methods generate a large number of plausible predictions and the challenge is to pick the 'best' one.

Coarse-Grained Protein Models

As computational power increases, many of the problems traditionally researched with coarse-grained protein models will be able to be attacked with all-atom models. In this light, it is interesting to consider the future of coarse-grained protein models and the types of problems they are likely to be solving. One promising area for CG models is the study of larger, even mesoscale biophysical systems (151), such as the adenylate kinase enzyme catalytic cycle (152).

Combining CG models with all-atom models in order to develop hybrid CG-AA models is an exciting idea (153). The aim is to use AA (or, in the future, even quantum mechanical) models for the parts of the system which are of particular interest or complexity, and then use a CG model for the more simple parts. The overall aim is to maintain the accuracy of all-atom models, whilst significantly reducing the computational expense. A review of the latest multiscale modelling frameworks can be found in (244).

6.2.3 β-Contact Prediction for Protein Structure and Protein-Protein Interaction Prediction

Since the publication of our β-contact prediction algorithm, Savojardo *et al.* have developed BCov, an alternative, unsupervised β-contact prediction algorithm (245). In their work they benchmark our algorithm on a new 1452 protein dataset and find that the accuracy of our approach matches that of BCov at the residue level. However, due to the more sophisticated way of predicting β-strand partners, BCov performs slightly better at the strand level.

This suggests it may be possible to improve our model to increase the accuracy of strand level predictions, and it may be useful to take into account the whole protein topology within the model. For example, currently, once the interacting strands have been chosen, the direction of interaction is independent for each pair, whereas this is not actually the case in proteins (198). It is also true that certain β-topologies are much more common than others (198) and it is possible our model could use this information effectively. A further project would be to develop a model which could jointly infer both the protein secondary structure and the β-contacts.

The direct information statistic has been used to study the co-evolution of inter-chain residues of interacting proteins (219). It would be an interesting project to see if the DI can be incorporated into a protein-protein interaction (PPI) prediction algorithm. Similar to the β-contact prediction case, if the protein MSAs are reasonably small, then the DI could be coupled to a PPI model, or used as one of the inputs to a scoring function aiming to rank possible docked PPI configurations. See (246), and references therein, for further details.

General Protein Structure Pipelines

It has previously been noted that contact prediction methods which rely on correlated mutation measures typically require a large number of homologous sequences, and therefore there is often a structure for one of these homologous proteins. If so, template methods, which are more accurate than *ab initio* techniques, can be used (213).

Therefore, it would also be interesting to test how well our model, and protein β-contact prediction algorithms in general, can be used within tertiary structure prediction pipelines. For example, is it preferable to know β-contacts rather than general protein contacts, and is it easier to predict them?

Different models incorporate a large variety of sources of information, for example β-sheet topologies, existing protein fragments, putative protein contacts, evolutionarily information, amino acid chemical properties and even early-stage experimental data. Combining these models in useful, yet statistically rigorous ways, is another important area of research, both for protein science and more general 'big data' problems.

6.3 Final Reflections

In this work we have developed and applied statistical methods to the study of outstanding research problems in computational structural biology. We have used the statistical machine learning technique, contrastive divergence, within a maximum likelihood framework to optimize the parameters of a coarse-grained protein model. We used the Bayesian inference algorithm nested sampling to explore the protein folding landscapes of this model. We combined a maximum entropy-derived correlated mutation measure with an empirical Bayes model in order to improve the prediction of protein β-contacts for use within tertiary structure prediction pipelines. Finally we implemented a novel phase space exploration method, Galilean sampling, in order to use nested sampling within a molecular dynamics framework.

Throughout the work we have benchmarked our new algorithms against existing approaches. We have compared nested sampling to parallel tempering and Galilean nested sampling to replica exchange molecular dynamics when calculating peptide heat capacities, and we have compared our protein β-contact prediction algorithm against existing state-of-the-art algorithms BetaPro and MLN.

We have used our algorithms to study specific biophysical problems of interest. We investigated the performance of the latest Amber force field by calculating both the free energy surface and the heat capacity curve of alanine dipeptide. We studied the behaviour of different functional forms when modelling atomic steric effects. Using energy landscape charts, we visualized the potential energy surface of three small proteins. We demonstrated the impressive nature of the maximum entropy-based correlated mutation measure, the direct information, using a large set of test proteins. Finally, drawing together our separate algorithms, we successfully folded two CASP targets using nested sampling and CRANKITE, using inputs from our β-contact prediction algorithm.

I look forward to seeing how the work in this thesis, and more generally, the application of novel statistical techniques to protein science, advance the field over the coming years.

Bibliography

[1] Chandler, D., 1987. Introduction to modern statistical mechanics. Oxford University Press, Oxford, UK.

[2] Roy, B. N., 2002. Fundamentals of classical and statistical thermodynamics. John Wiley & Sons, Hoboken, NJ, USA.

[3] Wales, D., 2003. Energy landscapes: Applications to clusters, biomolecules and glasses. Cambridge University Press, Cambridge, UK.

[4] Kadanoff, L. P., 2000. Statistical Physics Statics, Dynamics and Renormalization. World Scientific, Singapore.

[5] Finkelstein, A. V., and O. Ptitsyn, 2002. Protein physics: a course of lectures, Academic Press, San Diego, CA, USA, Chapter 8.

[6] Leach, A. R., 2001. Molecular modelling: principles and applications. Pearson Education, Harlow, UK.

[7] Alder, B. J., and T. Wainwright, 2004. Studies in molecular dynamics. I. General method. J. Chem. Phys. 31:459–466.

[8] Swope, W. C., H. C. Andersen, P. H. Berens, and K. R. Wilson, 1982. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. J. Chem. Phys. 76:637–649.

[9] Andersen, H. C., 1980. Molecular dynamics simulations at constant pressure and/or temperature. J. Chem. Phys. 72:2384–2393.

[10] Frenkel, D., and B. Smit, 2001. Understanding molecular simulation: from algorithms to applications, Volume 1. Academic Press, San Diego, CA, USA.

[11] Metropolis, N., 1987. The Beginning of the Monte Carlo Method. Los Alamos Science 15:125.

[12] Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

[13] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

[14] Raftery, A. E., M. A. Newton, J. M. Satagopan, and P. N. Krivitsky, 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Oxford University Press, Oxford, UK.

[15] Kirkpatrick, S., M. Vecchi, et al., 1983. Optimization by simulated annealing. Science 220:671–680.

[16] Marinari, E., and G. Parisi, 1992. Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. 19:451–458.

[17] Swendsen, R. H., and J.-S. Wang, 1986. Replica Monte Carlo simulation of spin glasses. Phys. Rev. Lett. 57:2607–2609.

[18] Ferrenberg, A. M., and R. H. Swendsen, 1989. Optimized Monte Carlo data analysis. Phys. Rev. Lett. 63:1195–1198.

[19] Sugita, Y., and Y. Okamoto, 1999. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 314:141–151.

[20] Katzgraber, H. G., S. Trebst, D. A. Huse, and M. Troyer, 2006. Feedback-optimized parallel tempering Monte Carlo. J. Stat. Mech. Theor. Exp. 2006:P03018.

[21] Earl, D. J., and M. W. Deem, 2005. Parallel tempering: Theory, applications, and new perspectives. Phys. Chem. Chem. Phys. 7:3910–3916.

[22] Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth, 1987. Hybrid Monte Carlo. Phys. Lett. B 195:216–222.

[23] Berg, B. A., and T. Neuhaus, 1992. Multicanonical ensemble: A new approach to simulate first-order phase transitions. Phys. Rev. Lett. 68:9–12.

[24] Wang, F., and D. P. Landau, 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys. Rev. Lett. 86:2050–2053.

[25] Kou, S., Q. Zhou, W. H. Wong, et al., 2006. Equi-energy sampler with applications in statistical inference and statistical mechanics. Ann. Stat. 34:1581–1619.

[26] Hesselbo, B., and R. B. Stinchcombe, 1995. Monte Carlo simulation and global optimization without parameters. Phys. Rev. Lett. 74:2151–2155.

[27] Neal, R. M., 2001. Annealed importance sampling. Stat. Comput. 11:125–139.

[28] Azevedo-Filho, A., and R. D. Shachter, 1994. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 28–36.

[29] Gelman, A., and X.-L. Meng, 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat. Science 163–185.

[30] Meng, X.-L., and W. H. Wong, 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica 6:831–860.

[31] Straatsma, T., and H. Berendsen, 1988. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. J. Chem. Phys. 89:5876–5886.

[32] Kirkwood, J. G., 1935. Statistical mechanics of fluid mixtures. J. Chem. Phys. 3:300–313.

[33] Torrie, G. M., and J. P. Valleau, 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J. Comput. Phys. 23:187–199.

[34] Kumar, S., J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, 1992. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. 13:1011–1021.

[35] Zwanzig, R. W., 1954. High-temperature equation of state by a perturbation method. I. nonpolar gases. J. Chem. Phys. 22:1420–1426.

[36] Kong, X., and C. L. Brooks III, 1996. λ-dynamics: A new approach to free energy calculations. J. Chem. Phys. 105:2414–2423.

[37] Frenkel, D., and A. J. Ladd, 1984. New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. J. Chem. Phys. 81:3188–3193.

[38] Skilling, J., 2004. Nested sampling. AIP Conf. Proc. 735:395–405.

[39] Skilling, J., 2006. Nested Sampling for General Bayesian Computation. J. Bayesian Anal. 1:833–859.

[40] Mukherjee, P., D. Parkinson, and A. R. Liddle, 2006. A nested sampling algorithm for cosmological model selection. Astrophys. J. Lett. 638:L51.

[41] Pullen, N., and R. J. Morris, 2014. Bayesian Model Comparison and Parameter Inference in Systems Biology Using Nested Sampling. PLOS ONE 9:e88419.

[42] Doğruel, M., T. A. Down, and T. J. Hubbard, 2008. NestedMICA as an ab initio protein motif discovery tool. BMC Bioinformatics 9:19–31.

[43] Pártay, L. B., A. P. Bartók, and G. Csányi, 2010. Efficient sampling of atomic configurational spaces. J. Phys. Chem. B 114:10502–10512.

[44] Lesk, A. M., 2004. Introduction to protein science: architecture, function, and genomics. Oxford University Press Oxford, UK.

[45] Kendrew, J. C., G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. Phillips, 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662–666.

[46] Perutz, M. F., M. Rossmann, A. F. Cullis, H. Muirhead, and G. Will, 1960. Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. Nature 185:416–422.

[47] Wüthrich, K., M. Billeter, and W. Braun, 1984. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. J. Mol. Biol. 180:715–740.

[48] Fasman, G. D., 1996. Circular dichroism and the conformational analysis of biomolecules. Springer, New York City, NY, USA.

[49] Giepmans, B. N., S. R. Adams, M. H. Ellisman, and R. Y. Tsien, 2006. The fluorescent toolbox for assessing protein location and function. Science Signaling 312:217–224.

[50] Lau, K. F., and K. A. Dill, 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22:3986–3997.

[51] Lindorff-Larsen, K., S. Piana, R. O. Dror, and D. E. Shaw, 2011. How Fast-Folding Proteins Fold. Science 334:517–520.

[52] Humphrey, W., A. Dalke, and K. Schulten, 1996. VMD – Visual Molecular Dynamics. J. Mol. Graphics 14:33–38.

[53] Linderstrøm-Lang, K. U., 1952. Lane Medical Lectures: proteins and enzymes, Volume 6. Stanford University Press, Stanford, CA, USA.

[54] Tanford, C., 1980. The Hydrophobic Effect: Formation of Micelles and Biological Membranes 2nd Ed. John Wiley & Sons, Hoboken, NJ, USA.

[55] Spolar, R. S., J.-H. Ha, and M. T. Record, 1989. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. Proc. Natl. Acad. Sci. USA 86:8382–8385.

[56] Honig, B. H., W. L. Hubbell, and R. F. Flewelling, 1986. Electrostatic interactions in membranes and proteins. Annu. Rev. Biophys. Biophys. Chem. 15:163–193.

[57] Roth, C. M., B. L. Neal, and A. M. Lenhoff, 1996. Van der Waals interactions involving proteins. Biophys. J. 70:977–987.

[58] Sigel, H., and R. B. Martin, 1982. Coordinating properties of the amide bond. Stability and structure of metal ion complexes of peptides and related ligands. Chem. Rev. 82:385–426.

[59] Thornton, J., 1981. Disulphide bridges in globular proteins. J. Mol. Biol. 151:261–287.

[60] Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia, 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540.

[61] Orengo, C. A., A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton, 1997. CATH–a hierarchic classification of protein domain structures. Structure 5:1093–1109.

[62] Anfinsen, C., 1973. Principles that govern the protein folding chains. Science 181:233–230.

[63] Dunker, A. K., J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, 2001. Intrinsically disordered protein. J. Mol. Graphics Modell. 19:26–59.

[64] Eliezer, D., 2009. Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 19:23–30.

[65] Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes, 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Struct. Funct. Bioinf. 21:167–195.

[66] Dill, K. A., and H. S. Chan, 1997. From Levinthal to pathways to funnels. Nature Struct. Biol. 4:10–19.

[67] Dinner, A. R., A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus, 2000. Understanding protein folding via free-energy surfaces from theory and experiment. Trends Biochem. Sci. 25:331–339.

[68] Onuchic, J. N., and P. G. Wolynes, 2004. Theory of protein folding. Curr. Opin. Struct. Biol. 14:70–75.

[69] Taverna, D. M., and R. A. Goldstein, 2002. Why are proteins marginally stable? Proteins: Struct. Funct. Bioinf. 46:105–109.

[70] Bryngelson, J. D., and P. G. Wolynes, 1989. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. 93:6902–6915.

[71] Baldwin, R. L., and G. D. Rose, 1999. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem. Sci. 24:26–33.

[72] Baldwin, R. L., and G. D. Rose, 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem. Sci. 24:77–83.

[73] Matouschek, A., J. T. Kellis, L. Serrano, and A. R. Fersht, 1989. Mapping the transition state and pathway of protein folding by protein engineering. Nature 340:122–126.

[74] Shakhnovich, E. I., and A. V. Finkelstein, 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. Biopolymers 28:1667–1680.

[75] Nojima, H., A. Ikai, T. Oshima, and H. Noda, 1977. Reversible thermal unfolding of thermostable phosphoglycerate kinase. Thermostability associated with mean zero enthalpy change. J. Mol. Biol. 116:429–442.

[76] Privalov, P., and N. Khechinashvili, 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. J. Mol. Biol. 86:665–684.

[77] Finkelstein, A. V., and O. Ptitsyn, 2002. Protein physics: a course of lectures., Academic Press, San Diego, CA, USA, Chapter 17–21.

[78] Piana, S., K. Lindorff-Larsen, and D. E. Shaw, 2013. Atomic-level description of ubiquitin folding. Proc. Natl. Acad. Sci. USA 110:5915–5920.

[79] Karanicolas, J., and C. L. Brooks, 2002. The origins of asymmetry in the folding transition states of protein L and protein G. Protein Sci. 11:2351–2361.

[80] Pauling, L., and R. B. Corey, 1951. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl. Acad. Sci. USA 37:729–740.

[81] Senn, H. M., and W. Thiel, 2009. QM/MM methods for biomolecular systems. Angew. Chem. Int. Ed. 48:1198–1229.

[82] Wang, J., R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, 2004. Development and testing of a general amber force field. J. Comput. Chem. 25:1157–1174.

[83] Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., 2009. CHARMM: the biomolecular simulation program. J. Comput. Chem. 30:1545–1614.

[84] Li, A., and V. Daggett, 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. Proc. Natl. Acad. Sci. USA 91:10430–10434.

[85] Mayor, U., C. M. Johnson, V. Daggett, and A. R. Fersht, 2000. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc. Natl. Acad. Sci. USA 97:13518–13522.

[86] Tozzini, V., 2010. Minimalist models for proteins: a comparative analysis. Q. Rev. Biophys. 43:333–371.

[87] Takada, S., 2012. Coarse-grained molecular simulations of large biomolecules. Curr. Opin. Struct. Biol. 22:130–137.

[88] Gō, N., 1983. Theoretical studies of protein folding. Annu. Rev. Biophys. Bioeng. 12:183–210.

[89] Takada, S., 1999. Gō-ing for the prediction of protein folding mechanisms. Proc. Natl. Acad. Sci. USA 96:11698–11700.

[90] Dudko, O. K., T. G. Graham, and R. B. Best, 2011. Locating the barrier for folding of single molecules under an external force. Phys. Rev. Lett. 107:208301.

[91] Tirion, M. M., 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 77:1905–1908.

[92] Bahar, I., A. R. Atilgan, and B. Erman, 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design 2:173–181.

[93] Suhre, K., and Y.-H. Sanejouand, 2004. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucl. Acids Res. 32:W610–W614.

[94] Wells, S., S. Menor, B. Hespenheide, and M. Thorpe, 2005. Constrained geometric simulation of diffusive motion in proteins. Phys. Biol. 2:S127–S136.

[95] Jacobs, D. J., A. J. Rader, L. A. Kuhn, and M. F. Thorpe, 2001. Protein flexibility predictions using graph theory. Proteins: Struct. Funct. Bioinf. 44:150–165.

[96] Nguyen, H. D., and C. K. Hall, 2004. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc. Natl. Acad. Sci. USA 101:16180–16185.

[97] Terakawa, T., and S. Takada, 2011. Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain. Biophys. J. 101:1450–1458.

[98] Monticelli, L., S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, 2008. The MARTINI coarse-grained force field: extension to proteins. J. Chem. Theory Comput. 4:819–834.

[99] Hardin, C., M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, 2000. Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. Proc. Natl. Acad. Sci. USA 97:14235–14240.

[100] Podtelezhnikov, A. A., and D. L. Wild, 2005. Exhaustive Metropolis Monte Carlo sampling and analysis of polyalanine conformations adopted under the influence of hydrogen bonds. Proteins: Struct. Funct. Bioinf. 61:94–104.

[101] Podtelezhnikov, A. A., and D. L. Wild, 2008. CRANKITE: A fast polypeptide backbone conformation sampler. Source Code Biol. Med. 3:1–7.

[102] Honeycutt, J., and D. Thirumalai, 1990. Metastability of the folded states of globular proteins. Proc. Natl. Acad. Sci. USA 87:3526–3529.

[103] Korkut, A., and W. A. Hendrickson, 2009. A force field for virtual atom molecular mechanics of proteins. Proc. Natl. Acad. Sci. USA 106:15667–15672.

[104] Becker, O. M., and M. Karplus, 1997. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. 106:1495–1517.

[105] Podtelezhnikov, A. A., Z. Ghahramani, and D. L. Wild, 2007. Learning about protein hydrogen bonding by minimizing contrastive divergence. Proteins: Struct. Funct. Bioinf. 66:588–599.

[106] Winther, O., and A. Krogh, 2004. Teaching computers to fold proteins. Phys. Rev. E 70:030903.

[107] Shell, M. S., 2008. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J. Chem. Phys. 129:144108.

[108] Duan, Y., C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. 24:1999–2012.

[109] MacKerell, A. D., D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102:3586–3616.

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

[111] Mullinax, J., and W. Noid, 2010. Recovering physical potentials from a model protein databank. Proc. Natl. Acad. Sci. USA 107:19867–19872.

[112] Maiorov, V. N., and G. M. Grippen, 1992. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888.

[113] Goldstein, R. A., Z. A. Luthey-Schulten, and P. G. Wolynes, 1992. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc. Natl. Acad. Sci. USA 89:9029–9033.

[114] Maupetit, J., P. Tuffery, and P. Derreumaux, 2007. A coarse-grained protein force field for folding and structure prediction. Proteins: Struct. Funct. Bioinf. 69:394–408.

[115] Carmichael, S. P., and M. S. Shell, 2012. A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. J. Phys. Chem. B 116:8383–8393.

[116] Podtelezhnikov, A. A., and D. L. Wild, 2009. Reconstruction and Stability of Secondary Structure Elements in the Context of Protein Structure Prediction. Biophys. J. 96:4399–4408.

[117] Hinton, G. E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14:1771–1800.

[118] Elofsson, A., S. M. Le Grand, and D. Eisenberg, 1995. Local moves: An efficient algorithm for simulation of protein folding. Proteins: Struct. Funct. Bioinf. 23:73–82.

[119] Simmerling, C., B. Strockbine, and A. E. Roitberg, 2002. All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. 124:11258–11259.

[120] van der Spoel, D., and M. M. Seibert, 2006. Protein folding kinetics and thermodynamics from atomistic simulations. Phys. Rev. Lett. 96:238102.

[121] Yang, S., J. N. Onuchic, A. E. García, and H. Levine, 2007. Folding time predictions from all-atom replica exchange simulations. J. Mol. Biol. 372:756–763.

[122] Jimenez-Cruz, C. A., and A. E. Garcia, 2013. Reconstructing the most probable folding transition path from replica exchange molecular dynamics simulations. J. Chem. Theory Comput. 9:3750–3755.

[123] Wu, X., and B. R. Brooks, 2003. Self-guided Langevin dynamics simulation method. Chem. Phys. Lett. 381:512–518.

[124] Wales, D. J., 2002. Discrete path sampling. Mol. Phys. 100:3285–3305.

[125] Evans, D. A., and D. J. Wales, 2004. Folding of the GB1 hairpin peptide from discrete path sampling. J. Chem. Phys. 121:1080–1090.

[126] Rapaport, D., 1979. Molecular dynamics study of a polymer chain in solution. J. Chem. Phys. 71:3299–3303.

[127] Dokholyan, N. V., S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, 1998. Discrete molecular dynamics studies of the folding of a protein-like model. Folding and Design 3:577–587.

[128] Brooks, B., and M. Karplus, 1983. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. USA 80:6571–6575.

[129] Piana, S., J. L. Klepeis, and D. E. Shaw, 2014. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr. Opin. Struct. Biol. 24:98–105.

[130] Hu, H., M. Elstner, and J. Hermans, 2003. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. Proteins: Struct. Funct. Genet. 50:451–463.

[131] Mu, Y., D. S. Kosov, and G. Stock, 2003. Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. J. Phys. Chem. B 107:5064–5073.

[132] Shen, M.-Y., and K. F. Freed, 2002. Long time dynamics of met-enkephalin: comparison of explicit and implicit solvent models. Biophys. J. 82:1791–1808.

[133] Mehta, A. D., M. Rief, J. A. Spudich, D. A. Smith, and R. M. Simmons, 1999. Single-molecule biomechanics with optical methods. Science 283:1689–1695.

[134] Yeh, I.-C., M. S. Lee, and M. A. Olson, 2008. Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. J. Phys. Chem. B 112:15064–15073.

[135] Lee, M. S., and M. A. Olson, 2011. Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. J. Chem. Phys. 134:244111.

[136] Shea, J.-E., J. N. Onuchic, and C. L. Brooks, 2002. Probing the folding free energy landscape of the src-SH3 protein domain. Proc. Natl. Acad. Sci. USA 99:16064–16068.

[137] Zhou, R., 2003. Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins: Struct. Funct. Bioinf. 53:148–161.

[138] Becker, O. M., and M. Karplus, 1997. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. 106:1495–1517.

[139] Wales, D. J., and T. V. Bogdan, 2006. Potential energy and free energy landscapes. J. Phys. Chem. B 110:20765–20776.

[140] Koga, N., and S. Takada, 2006. Folding-based molecular simulations reveal mechanisms of the rotary motor F1–ATPase. Proc. Natl. Acad. Sci. USA 103:5367–5372.

[141] Fawzi, N. L., Y. Okabe, E.-H. Yap, and T. Head-Gordon, 2007. Determining the Critical Nucleus and Mechanism of Fibril Elongation of the Alzheimers Aβ_{1-40} Peptide. J. Mol. Biol. 365:535–550.

[142] Levinthal, C., S. J. Wodak, P. Kahn, and A. K. Dadivanian, 1975. Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. Proc. Natl. Acad. Sci. USA 72:1330–1334.

[143] Dominguez, C., R. Boelens, and A. M. Bonvin, 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. 125:1731–1737.

[144] Bonvin, A., 2013. Coming to peace with protein complexes? 5th CAPRI evaluation meeting, April 17–19th 2013–Utrecht. Proteins: Struct. Funct. Bioinf. 81:2073–2074.

[145] Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368.

[146] Shimaoka, M., J. M. Shifman, H. Jing, J. Takagi, S. L. Mayo, and T. A. Springer, 2000. Computational design of an integrin I domain stabilized in the open high affinity conformation. Nature Struct. Mol. Biol. 7:674–678.

[147] Shifman, J. M., and S. L. Mayo, 2003. Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc. Natl. Acad. Sci. USA 100:13274–13279.

[148] Allen, B. D., A. Nisthal, and S. L. Mayo, 2010. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc. Natl. Acad. Sci. USA 107:19838–19843.

[149] Yang, S., L. Blachowicz, L. Makowski, and B. Roux, 2010. Multidomain assembled states of Hck tyrosine kinase in solution. Proc. Natl. Acad. Sci. USA 107:15757–15762.

[150] Chebaro, Y., X. Dong, R. Laghaei, P. Derreumaux, and N. Mousseau, 2008. Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent. J. Phys. Chem. B 113:267–274.

[151] Kapral, R., 2008. Multiparticle collision dynamics: simulation of complex systems on mesoscales. Adv. Chem. Phys. 140:89.

[152] Echeverria, C., Y. Togashi, A. S. Mikhailov, and R. Kapral, 2011. A mesoscopic model for protein enzymatic dynamics in solution. Phys. Chem. Chem. Phys. 13:10527–10537.

[153] Li, W., and S. Takada, 2010. Characterizing Protein Energy Landscape by Self-Learning Multiscale Simulations: Application to a Designed β-Hairpin. Biophys. J. 99:3029–3037.

[154] Kotelchuck, D., and H. Scheraga, 1969. The influence of short-range interactions on protein conformation, II. A model for predicting the α-helical regions of proteins. Proc. Natl. Acad. Sci. USA 62:14–21.

[155] Zhang, Y., 2008. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 18:342–348.

[156] Zhang, Y., 2009. Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 19:145–155.

[157] The UniProt Consortium, 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucl. Acids Res. 41:D43–D47.

[158] Bernstein, F., T. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, 1977. Protein Data Bank – Computer-based Archival File for Macromolecular Structures. J. Mol. Biol. 112:535–542.

[159] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, 2000. The protein data bank. Nucl. Acids Res. 28:235–242.

[160] Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25:3389–3402.

[161] Söding, J., 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.

[162] Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

[163] Zhu, J., H. Fan, X. Periole, B. Honig, and A. E. Mark, 2008. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins: Struct. Funct. Bioinf. 72:1171–1188.

[164] Hildebrand, A., M. Remmert, A. Biegert, and J. Soeding, 2009. Fast and accurate automatic structure prediction with HHpred. Proteins: Struct. Funct. Bioinf. 77:128–132.

[165] Xu, D., J. Zhang, A. Roy, and Y. Zhang, 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins: Struct. Funct. Bioinf. 79:147–160.

[166] Mariani, V., F. Kiefer, T. Schmidt, J. Haas, and T. Schwede, 2011. Assessment of template based protein structure predictions in CASP9. Proteins: Struct. Funct. Bioinf. 79:37–58.

[167] Bowie, J. U., R. Luthy, and D. Eisenberg, 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170.

[168] Jones, D. T., W. Taylort, and J. M. Thornton, 1992. A new approach to protein fold recognition. Nature 358:86–89.

[169] Bennett-Lovsey, R. M., A. D. Herbert, M. J. Sternberg, and L. A. Kelley, 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins: Struct. Funct. Bioinf. 70:611–625.

[170] Kelley, L. A., and M. J. Sternberg, 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4:363–371.

[171] Wolf, Y. I., N. V. Grishin, and E. V. Koonin, 2000. Estimating the number of protein folds and families from complete genome data. J. Mol. Biol. 299:897–905.

[172] Kinch, L., S. Y. Shi, Q. Cong, H. Cheng, Y. Liao, and N. V. Grishin, 2011. CASP9 assessment of free modeling target predictions. Proteins: Struct. Funct. Bioinf. 79:59–73.

[173] Tai, C.-H., H. Bai, T. J. Taylor, and B. Lee, 2014. Assessment of template-free modeling in CASP10 and ROLL. Proteins: Struct. Funct. Bioinf. 82:57–83.

[174] Simons, K., R. Bonneau, I. Ruczinski, and D. Baker, 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Struct. Funct. Bioinf. 171–176. 3rd Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP3), Pacific Grove, CA.

[175] Raman, S., R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker, 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins: Struct. Funct. Bioinf. 77:89–99.

[176] Khatib, F., S. Cooper, M. D. Tyka, K. Xu, I. Makedon, Z. Popovic, D. Baker, and Foldit Players, 2011. Algorithm discovery by protein folding game players. Proc. Natl. Acad. Sci. USA 108:18949–18953.

[177] Rost, B., 2001. Review: Protein secondary structure prediction continues to rise. J. Struct. Biol. 134:204–218.

[178] Qian, N., and T. J. Sejnowski, 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884.

[179] Pollastri, G., D. Przybylski, B. Rost, and P. Baldi, 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Struct. Funct. Genet. 47:228–235.

[180] Hua, S., and Z. Sun, 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. J. Mol. Biol. 308:397–407.

[181] Schmidler, S. C., J. S. Liu, and D. L. Brutlag, 2000. Bayesian segmentation of protein secondary structure. J. Comp. Biol. 7:233–248.

[182] Chu, W., Z. Ghahramani, A. Podtelezhnikov, and D. Wild, 2006. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. IEEE/ACM Trans. Comput. Biol. Bioinformatics 3:98–113.

[183] Cuff, J., and G. Barton, 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Struct. Funct. Bioinf. 40:502–511.

[184] Aydin, Z., Y. Altunbasak, and M. Borodovsky, 2006. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. BMC Bioinformatics 7:178.

[185] Yuan, Z., K. Burrage, and J. Mattick, 2002. Prediction of protein solvent accessibility using support vector machines. Proteins: Struct. Funct. Genet. 48:566–570.

[186] Ferre, F., and P. Clote, 2005. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 21:2336–2346.

[187] Monastyrskyy, B., D. D'Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych, 2014. Evaluation of residue-residue contact prediction in CASP10. Proteins: Struct. Funct. Bioinf. 82:138–153.

[188] Phillips, D., 1970. The development of crystallographic enzymology. In Biochem. Soc. Symp. Academic Press., San Diego, CA, USA, Volume 30, 11–28.

[189] Fariselli, P., and R. Casadio, 1999. A neural network based predictor of residue contacts in proteins. Protein Eng. 12:15–21.

[190] Di Lena, P., K. Nagata, and P. Baldi, 2012. Deep architectures for protein contact map prediction. Bioinformatics 28:2449–2457.

[191] Cheng, J., and P. Baldi, 2007. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8:113.

[192] Li, Y., Y. Fang, and J. Fang, 2011. Predicting residue–residue contacts using random forest models. Bioinformatics 27:3379–3384.

[193] Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, 2011. Protein 3D structure computed from evolutionary sequence variation. PLOS ONE 6:e28766.

[194] Vendruscolo, M., E. Kussell, and E. Domany, 1997. Recovery of protein structure from contact maps. Folding and Design 2:295–306.

[195] Sathyapriya, R., J. M. Duarte, H. Stehr, I. Filippis, and M. Lappe, 2009. Defining an essence of structure determining residue contacts in proteins. PLoS Comput. Biol. 5:e1000584.

[196] Duarte, J. M., R. Sathyapriya, H. Stehr, I. Filippis, and M. Lappe, 2010. Optimal contact definition for reconstruction of contact maps. BMC Bioinformatics 11:283.

[197] Burkoff, N. S., C. Várnai, and D. L. Wild, 2013. Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. Bioinformatics 29:580–587.

[198] Ruczinski, I., C. Kooperberg, R. Bonneau, and D. Baker, 2002. Distributions of beta sheets in proteins with application to structure prediction. Proteins: Struct. Funct. Genet. 48:85–97.

[199] Kortemme, T., M. Ramırez-Alvarado, and L. Serrano, 1998. Design of a 20-amino acid, three-stranded β-sheet protein. Science 281:253–256.

[200] Smith, C. K., and L. Regan, 1995. Guidelines for protein design: The energetics of β-sheet side chain interactions. Science 270:980–980.

[201] Mandel-Gutfreund, Y., S. M. Zaremba, and L. M. Gregoret, 2001. Contributions of residue pairing to β-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel β-strands. J. Mol. Biol. 305:1145–1159.

[202] Merkel, J. S., and L. Regan, 2000. Modulating protein folding rates in vivo and in vitro by side-chain interactions between the parallel β strands of green fluorescent protein. J. Biol. Chem. 275:29200–29206.

[203] Cheng, J., and P. Baldi, 2005. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics 21:i75–i84.

[204] Hubbard, T. J., 1994. Use of β-strand interaction pseudo-potentials in protein structure prediction and modelling. In System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on. IEEE, Volume 5, 336–344.

[205] Steward, R. E., and J. M. Thornton, 2002. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. Proteins: Struct. Funct. Bioinf. 48:178–191.

[206] Aydin, Z., Y. Altunbasak, and H. Erdogan, 2011. Bayesian models and algorithms for protein β-sheet prediction. IEEE/ACM Trans. Comput. Biol. Bioinformatics 8:395–409.

[207] Lippi, M., and P. Frasconi, 2009. Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics 25:2326–2333.

[208] Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32:1792–1797.

[209] Notredame, C., D. G. Higgins, and J. Heringa, 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

[210] Göbel, U., C. Sander, R. Schneider, and A. Valencia, 1994. Correlated mutations and residue contacts in proteins. Proteins: Struct. Funct. Bioinf. 18:309–317.

[211] Olmea, O., and A. Valencia, 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding and Design 2:S25–S32.

[212] de Juan, D., F. Pazos, and A. Valencia, 2013. Emerging methods in protein co-evolution. Nat. Rev. Genet. 14:249–261.

[213] Taylor, W. R., R. S. Hamilton, and M. I. Sadowski, 2013. Prediction of contacts from correlated sequence substitutions. Curr. Opin. Struct. Biol. 23:473–479.

[214] Dekker, J. P., A. Fodor, R. W. Aldrich, and G. Yellen, 2004. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 20:1565–1572.

[215] Martin, L., G. B. Gloor, S. Dunn, and L. M. Wahl, 2005. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116–4124.

[216] Dunn, S. D., L. M. Wahl, and G. B. Gloor, 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340.

[217] Lee, B.-C., and D. Kim, 2009. A new method for revealing correlated mutations under the structural and functional constraints in proteins. Bioinformatics 25:2506–2513.

[218] Lapedes, A. S., B. G. Giraud, L. Liu, and G. D. Stormo, 1999. Correlated mutations in models of protein sequences: phylogenetic and structural effects, Institute of Mathematical Statistics, Ann Arbor, MI, USA, p236–256.

[219] Weigt, M., R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, 2009. Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. USA 106:67–72.

[220] Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149:1607–1621.

[221] Jaynes, E. T., 2003. Probability Theory: The Logic of Science. CUP, Cambridge, UK.

[222] Cocco, S., S. Leibler, and R. Monasson, 2009. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. Proc. Natl. Acad. Sci. USA 106:14058–14062.

[223] Burger, L., and E. van Nimwegen, 2010. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput. Biol. 6:e1000633.

[224] Skilling, J., 2012. Bayesian computation in big spaces-nested sampling and Galilean Monte Carlo. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Publishing, Melville, NY, USA, Volume 1443, 145–156.

[225] Brewer, B. J., L. B. Pártay, and G. Csányi, 2011. Diffusive nested sampling. Statistics and Computing 21:649–656.

[226] Shaw, R., M. Bridges, and M. Hobson, 2007. Clustered nested sampling: efficient Bayesian inference for cosmology. arXiv preprint astro-ph/0701867 .

[227] Abagyan, R., M. Totrov, and D. Kuznetsov, 1994. ICMa new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. 15:488–506.

[228] Yeh, I.-C., and A. Wallqvist, 2009. Structure and dynamics of end-to-end loop formation of the penta-peptide Cys-Ala-Gly-Gln-Trp in implicit solvents. J. Phys. Chem. B 113:12382–12390.

[229] Hamelberg, D., J. Mongan, and J. A. McCammon, 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. 120:11919–11929.

[230] Hansmann, U. H., Y. Okamoto, and F. Eisenmenger, 1996. Molecular dynamics, Langevin and hydrid Monte Carlo simulations in a multicanonical ensemble. Chem. Phys. Lett. 259:321–330.

[231] Berkholz, D. S., C. M. Driggers, M. V. Shapovalov, R. L. Dunbrack, and P. A. Karplus, 2012. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. Proc. Natl. Acad. Sci. USA 109:449–453.

[232] Irbäck, A., and S. Mohanty, 2006. PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J. Comput. Chem. 27:1548–1555.

[233] Mohanty, S., J. H. Meinke, and O. Zimmermann, 2013. Folding of Top7 in unbiased all-atom Monte Carlo simulations. Proteins: Struct. Funct. and Bioinf. 81:1446–1456.

[234] Lane, T. J., D. Shukla, K. A. Beauchamp, and V. S. Pande, 2013. To milliseconds and beyond: challenges in the simulation of protein folding. Curr. Opin. Struct. Biol. 23:58–65.

[235] Götz, A. W., M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, 2012. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. 8:1542–1555.

[236] Salomon-Ferrer, R., A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, 2013. Routine microsecond molecular dynamics simulations with Amber on GPUs. 2. Explicit solvent particle mesh Ewald. J. Chem. Theory Comput. 9:3878–3888.

[237] Shaw, D. E., M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, et al., 2008. Anton, a special-purpose machine for molecular dynamics simulation. Comm. ACM 51:91–97.

[238] Baker, C. M., and R. B. Best, 2014. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. WIREs Comput. Mol. Sci. 4:182–198.

[239] Ozenne, V., R. Schneider, M. Yao, J.-r. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge, 2012. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. J. Am. Chem. Soc. 134:15138–15148.

[240] Chodera, J. D., D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, 2011. Alchemical free energy methods for drug discovery: progress and challenges. Curr. Opin. Struct. Biol. 21:150–160.

[241] Hamelryck, T., M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, and J. Ferkinghoff-Borg, 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLOS ONE 5:e13714.

[242] Bottaro, S., K. Lindorff-Larsen, and R. B. Best, 2013. Variational Optimization of an All-Atom Implicit Solvent Force Field To Match Explicit Solvent Simulation Data. J. Chem. Theory Comput. 9:5641–5652.

[243] Kleinman, C. L., N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot, 2006. A maximum likelihood framework for protein design. BMC Bioinformatics 7:326.

[244] Zhou, H.-X., 2014. Theoretical frameworks for multiscale modeling and simulation. Curr. Opin. Struct. Biol. 25:67–76.

[245] Savojardo, C., P. Fariselli, P. L. Martelli, and R. Casadio, 2013. BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. Bioinformatics 29:3151–3157.

[246] Madaoui, H., and R. Guerois, 2008. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. Proc. Natl. Acad. Sci. USA 105:7708–7713.

Appendix

In Chapter 5 we sample Galilean velocities $\mathbf{v} = \mathbf{S}\mathbf{r}$ where $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{S}$ is proportional to the identity matrix. Following the MD implementation in Amber, we use Cartesian co-ordinates to describe the atomic positions, and hence we name this algorithm 'Cartesian isotropic' (CI).

As mentioned in Chapter 6, it is possible to choose a different 'semimetric' $\mathbf{S}$, in order to improve Galilean exploration. The reflection formula $\mathbf{v}' = \mathbf{v} - 2\mathbf{n}(\mathbf{n}^t\mathbf{v})$ is then replaced by

$$\mathbf{v}' = \mathbf{v} - 2\mathbf{S}\mathbf{S}^t\mathbf{n}\frac{\mathbf{n}^t\mathbf{v}}{\mathbf{n}^t\mathbf{S}\mathbf{S}^t\mathbf{n}},$$

in order to maintain detailed balance (224).

In Chapter 6 we claim that it will be essential to use an appropriate semimetric when using Galilean nested sampling with larger molecular systems. This is because in molecular systems, certain degrees of freedom, such as the stretching of covalent bonds, are very highly constrained, whereas others, such as the dihedral angles ϕ and ψ in proteins, are not very constrained at all. The magnitude of velocities in the highly constrained directions should be smaller than those in other directions.

Skilling suggests using $\mathbf{S} \approx (-\nabla\mathbf{F})^{-1/2}$, where $\mathbf{F}$ are the forces $(-\nabla E)$ at a preferred configuration (224). This choice of semimetric takes into account the curvature of the space. However, its calculation could be expensive and the choice of preferred configuration (especially using Cartesian co-ordinates) could be a difficult one to make. In this Appendix we briefly show how even a simple well-chosen semimetric can improve the Galilean exploration of the PESs of the small molecules ethane and alanine dipeptide.

Ethane

We first sample the PES of ethane (C_2H_6) using the GAFF Amber energy function (82). In molecular systems, including ethane, it is generally the case that covalent bond lengths are more constrained than valence angles, which themselves are more constrained than dihedral angles. We will use this information to devise a simple semimetric, and therefore we find it useful to describe the 18 dimensional PES using internal co-ordinates. Specifically, we parameterize the PES using $\{r, \mathbb{B}, \mathbb{A}, \mathbb{T}\}$, where r is the C-C bond distance, $\mathbb{B} = \{b_1, \ldots, b_6\}$ are the 6 C-H bond lengths, $\mathbb{A} = \{a_1, \ldots, a_6\}$ are the 6 H-C-C valence angles and $\mathbb{T} = \{t_1, \ldots, t_5\}$ are 5 of the H-C-C-H dihedral (torsion) angles. See Figure A1 for further details. The potential energy term is given by

$$E(\{r, \mathbb{B}, \mathbb{A}, \mathbb{T}\}) = K_r(r - r_{\text{eq}})^2 + \sum_{b \in \mathbb{B}} K_b(b - b_{\text{eq}})^2 + \sum_{a \in \mathbb{A}} K_a(a - a_{\text{eq}})^2 + \frac{1}{2}\sum_{t \in \mathbb{T}} V_3(1 + \cos 3t) + E_{\text{other}}(\{r, \mathbb{B}, \mathbb{A}, \mathbb{T}\})$$

where r_{eq}, b_{eq} and a_{eq} are equilibrium constants (whose values are unimportant here) and $K_r = 303.1$, $K_b = 337.3$, $K_a = 46.37$ and $V_3 = 0.15$ are force constants (values given in standard Amber units). The

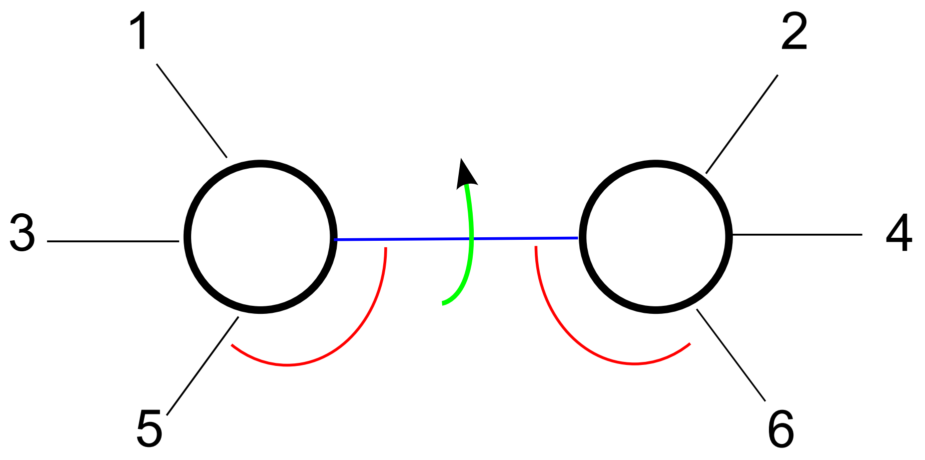


Figure A1: A molecule of ethane. Two carbon atoms (circles) covalently bonded to each other and to three hydrogen atoms (numbered 1 to 6). r is the C-C bond length (blue line), $b_1, \ldots, b_6$ are the 6 C-H bond lengths (black lines), $a_1, \ldots a_6$ are the 6 H-C-C valence angles (a_5 and a_6 are marked by red lines) and $t_1, \ldots, t_5$ are dihedral angles H-C-C-H, for different hydrogen atoms, marked by the green arrow.

E_{other} term contains all other energy terms. This includes potentials for the 6 valence and 5 dihedral angles whose values can be derived from $\{r, \mathbb{B}, \mathbb{A}, \mathbb{T}\}$ and Lennard Jones potentials between pairs of hydrogens bonded to different carbon atoms.[1]

Looking at the values of the force constants suggests a good choice of semimetric could be a simple diagonal semimetric $\mathbf{S} = \text{diag}(s_1, s_2, \ldots, s_{18})$, with s_i small when associated with bond length dimensions, larger for those associated with valence angle dimensions and even larger for the s_i for dihedral angle dimensions. Specifically, for bond length and valence angles we choose $s = (2K)^{-1/2}$, for the appropriate force constant K as this corresponds to Skilling's choice of semimetric, if E_{other} is ignored.

The functional form of the dihedral potential is more complex and we choose s_i to be 8.0 for dihedral angle dimensions.[2] This can be compared to ≈ 0.14 for valence angle dimensions. We name this semimetric 'internal anisotropic' (IA).

Comparison of semimetrics We briefly illustrate the behaviour of Galilean exploration of the PES of ethane using both the CI and IA semimetrics. These results are early proof-of-concept work and we expect to carry out further, more rigorous analysis in the future. For both semimetrics we set $\theta = 0.2$ (the amount of randomization of the velocity each step) and tune the timestep to keep a mean free path (average number of successful steps before leaving the acceptable region) of 5. This ensures the same number of force evaluations for both CI and IA.

The dihedral angle potential has 3 minima (±60°,180°), and, with a low nested sampling energy level, the dihedral angles of ethane cannot escape one of the minima; the molecule cannot cross the 'ethane barrier'. For this work we choose the current nested sampling energy level to be 10 kcal/mol, which is sufficiently high to keep the acceptable region connected.

Figure A2 shows the H-C-C-H dihedral angle (for two chosen hydrogen atoms) for eight CI trajectories (top left) and eight IA trajectories (top right). All 16 trajectories start from the same conformation and it is clearly much easier for the IA trajectories to cross the ethane barrier, whose location is marked

[1] Due to the geometry of ethane, the LJ functions are not especially important.

[2] This choice is somewhat arbitrary, but sufficient to illustrate the general point of the Appendix, which is that the choice of semimetric is very important.

with solid horizontal lines. This demonstrates how a sensible choice of semimetric can improve the exploration of the acceptable region, whilst, in the limit, sample it *uniformly*. Traces of the potential energy throughout a CI trajectory (bottom left) and an IA trajectory (bottom right) are also shown. For completeness, Figure A3 shows how a C-H bond distance (top) and a H-C-C valence angle (bottom) vary throughout CI (left) and IA (right) Galilean trajectories.

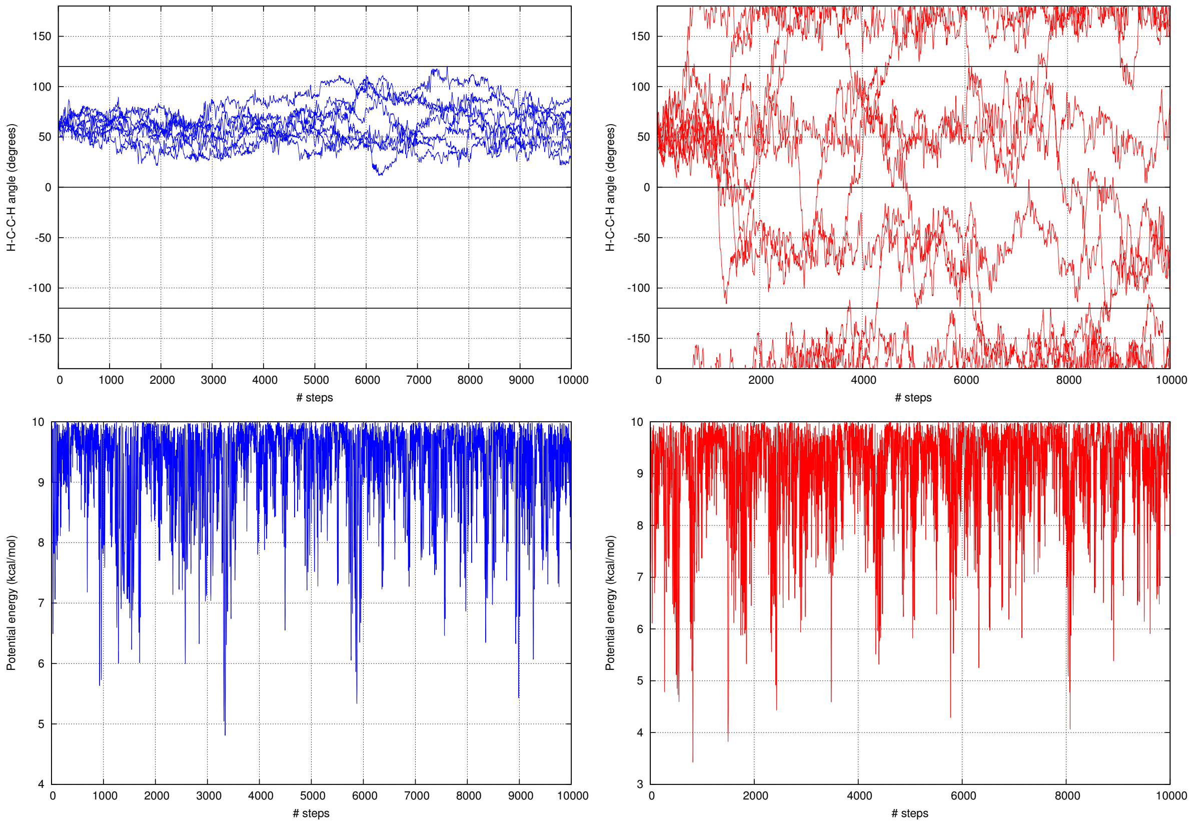


Figure A2: The value of an H-C-C-H dihedral angle for eight CI Galilean trajectories (top left) and eight IA trajectories (top right), with the current nested sampling energy level chosen to be 10 kcal/mol and all trajectories starting from the same initial configuration. It is clearly much easier for the IA trajectories to cross the ethane barrier, whose location is marked with solid horizontal lines. Traces of the potential energy throughout a CI trajectory (bottom left) and an IA trajectory (bottom right) are also shown.

Alanine dipeptide

A similar comparison of semimetrics can be made when sampling the PES of alanine dipeptide *in vacuo*. We use the same model and force field as described in Chapter 5, and in that chapter we used the Cartesian isotropic (CI) semimetric. Analogously to ethane, we can describe the 60 degrees of freedom of alanine dipeptide by 21 bond lengths, 20 valence angles and 19 dihedral angles. We can also define an internal anisotropic (IA) semimetric, by taking **S** to be diagonal with $(2K)^{-1/2}$ (for appropriate force constants K) for the bond length and valence angle dimensions; $s_i = 2.5$ for dihedral angles which keep the main chain atoms near planar; and $s_i = 10$ for the other dihedral angles including ϕ and ψ.

Figure A4 shows the trace of the (ϕ,ψ) angles on the Ramachandran plot for a single 10^5 step Galilean trajectory using the CI (left) and IA (right) semimetrics. The nested sampling energy level was chosen to be 100 kcal/mol, the mean free path was set to 3.5 and $\theta = 0.2$. It appears the IA is far superior at exploring the available region.

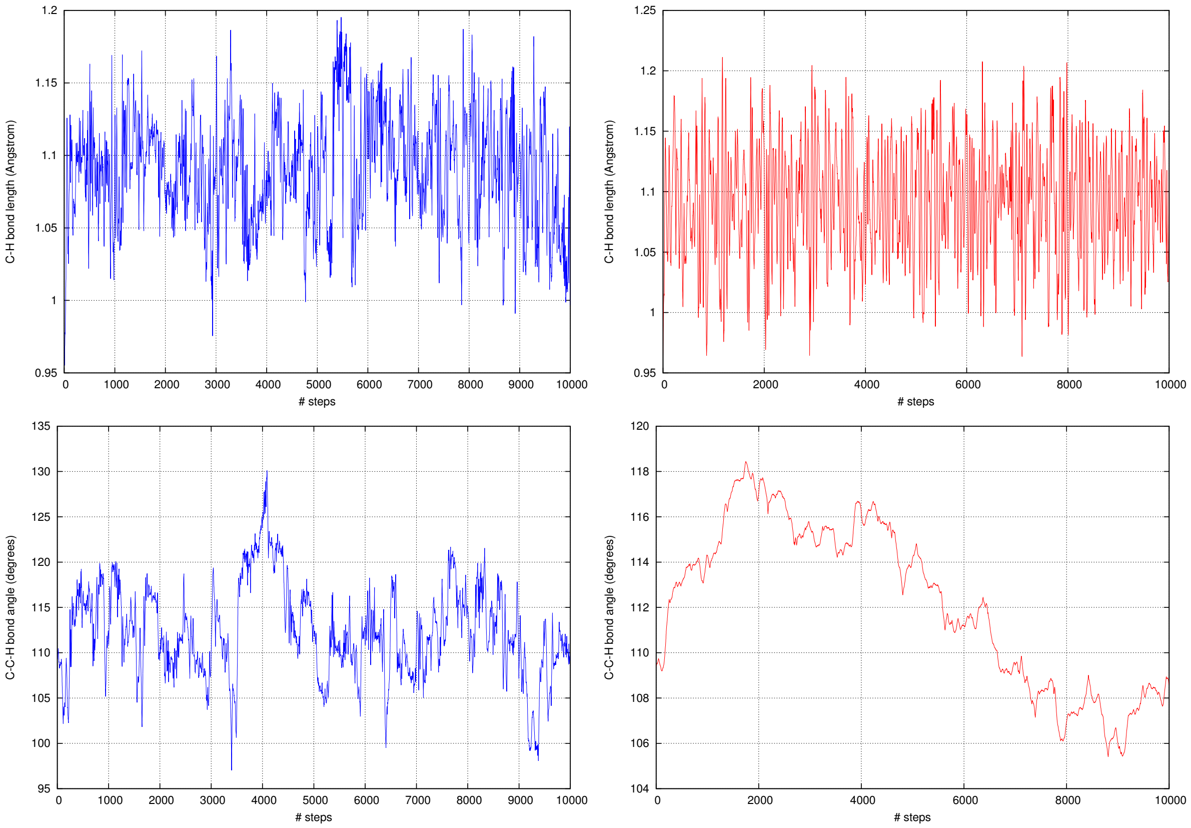


Figure A3: A C-H bond distance (top) and a H-C-C valence angle (bottom) trace plots throughout a CI (left) and IA (right) Galilean trajectory. The equilibrium bond length is 1.09Å and the equilibrium valence angle constant is 110°.

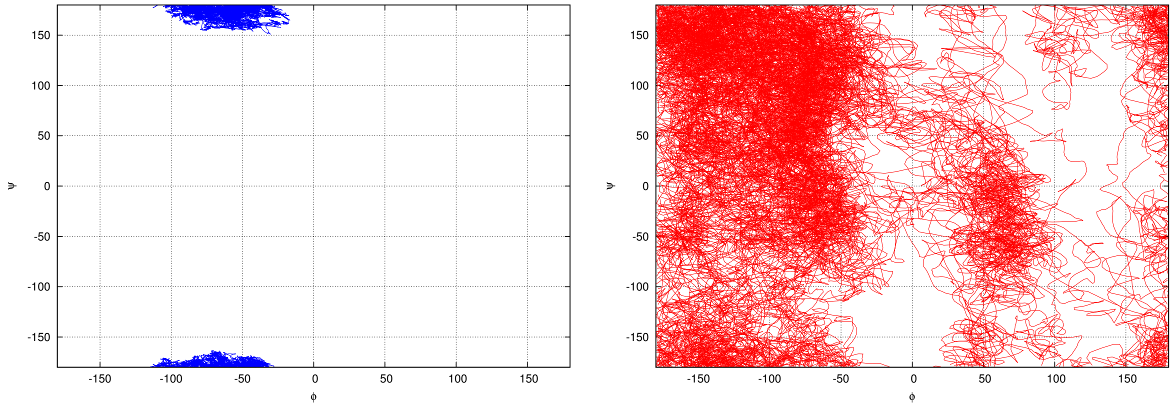


Figure A4: The trace of the (ϕ,ψ) angles on the Ramachandran plot for a single 10^5 step Galilean trajectory using the CI (left) and IA (right) semimetrics. The nested sampling energy level was chosen to be 100 kcal/mol, the mean free path was set to 3.5 and $\theta = 0.2$. It appears the IA is far superior at exploring the available region.

Figure A5 shows the empirical marginal (ϕ,ψ) probability distributions for sampling uniformly over the conformations with energy less than 100 kcal/mol using the CI metric (left) and the IA metric (right). The distributions were generated by running 8 independent trajectories for each semimetric (all starting from the same configuration) and discarding the first 10% of the trajectories for equilibration. A Gaussian filter was applied to smooth the data. The IA semimetric clearly explores the space more efficiently. It is interesting to compare Figure A5 to Figure S1 in Chapter 5. However, that figure was generated using a few orders of magnitude more force evaluations and used a very large number of short trajectories over a few nested sampling iterations, rather than a few long trajectories.

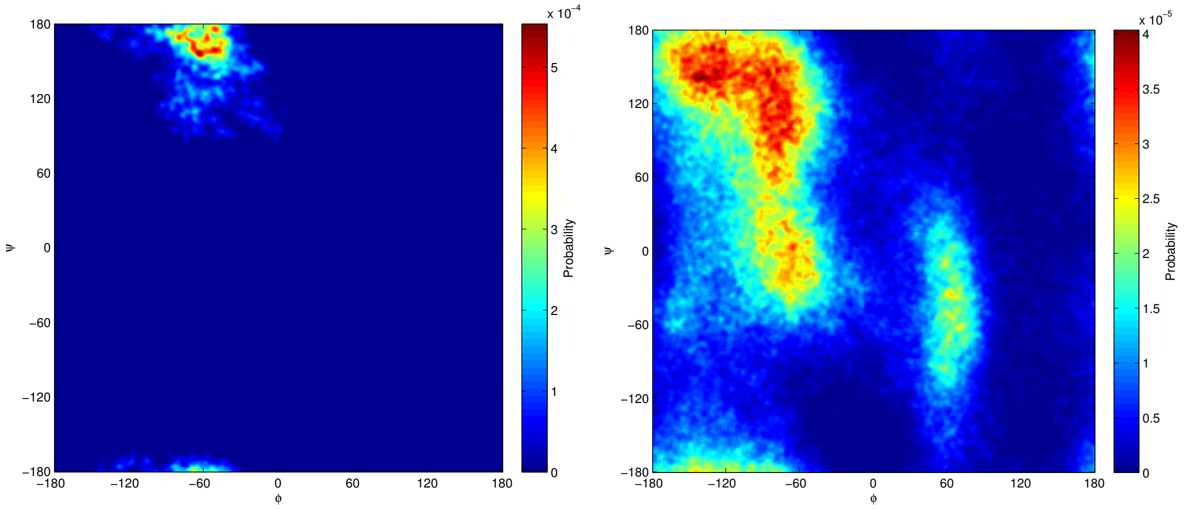


Figure A5: The empirical marginal (ϕ,ψ) probability distributions for sampling uniformly over the conformations with energy less than 100 kcal/mol using the CI metric (left) and the IA metric (right). The distributions were generated by running 8 independent trajectories for each semimetric, (all starting from the same configuration) and discarding the first 10% of the trajectories for equilibration. A Gaussian filter was applied to smooth the data.

Conclusion

In this Appendix we have shown that a simple, well-chosen semimetric can significantly improve the Galilean exploration of molecular systems. Unlike ethane, in larger molecules, E_{other} is significantly more complicated; it includes a larger number of derived valence and dihedral angle potentials along with improper dihedral angle potentials, significant LJ potentials, Coulombic point charge potentials and possibly generalized Born potentials for implicitly solvated systems. In these cases, the IA semimetric may perform less well. Further work will be required to compare the simple IA semimetric with the semimetric $\mathbf{S} \approx (-\nabla \mathbf{F})^{-1/2}$ suggested by Skilling.

From a practical point of view, Amber is not designed for internal co-ordinate molecular dynamics, and the preliminary results shown here involve a lot of highly inefficient converting between co-ordinate systems. This approach will definitely not scale to larger systems. Unlike the semimetric $\mathbf{S} \approx (-\nabla \mathbf{F})^{-1/2}$, which is independent of the co-ordinate system used, in order to use the IA semimetric with Amber efficiently, it is likely that large scale changes would need to be implemented. Therefore it may be necessary to use an alternative molecular dynamics package, for example the internal co-ordinate dynamics package of Abagyan *et al.* (227).