Computational Methods for Conformational Sampling of Biomolecules

Bottaro, Sandro; Ferkinghoff-Borg, Jesper; Lindorff-Larsen, Kresten

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Computational Methods for Conformational Sampling of Biomolecules

Sandro Bottaro

Technical University of Denmark

A thesis submitted for the degree of *Philosophiæ Doctor (PhD)*

April 2012

Supervisors: Jesper Ferkinghoff-Borg
Kresten Lindorff-Larsen

Preface

This dissertation is a summary of research carried out between April 2009 and April 2012 as a PhD candidate at the Technical University of Denmark. Part of the project was carried out as a visiting student at the Department of Biology, University of Copenhagen and at the Department of Chemistry, University of Cambridge, to which I hereby extend my gratitude. I also acknowledge Radiometer for financial support.

The central idea behind my studies is the development and use of methodologies for molecular simulations of biomolecules. In this thesis, three distinct but connected aspects of this problem are considered. First, the design and the assessment of a Monte Carlo method for conformational sampling of macromolecules. Secondly, the development of a model describing the water environment, and finally the applications of these techniques to biological problems.

An introduction to the central topics of the thesis, together with an overview of the literature, is given in the first chapter. The results of my research are presented in chapter 2, which constitutes the core of the dissertation. Concluding remarks and possible directions for future work are presented in the last chapter.

Sandro Bottaro

Copenhagen, April 2012

Abstract

Proteins play a fundamental role in virtually every process within living organisms. For example, some proteins act as enzymes, catalyzing a wide range of reactions necessary for life, others mediate the cell interaction with the surrounding environment and still others have regulatory functions. Recent studies have demonstrated that the specificity of the biological function and activity of proteins is intimately linked to their structural and dynamical properties. In principle, these properties can be calculated using computational techniques. However, most structural transitions of biological relevance occur on time-scales inaccessible to current methodologies due to prohibitive computational costs. In this dissertation I present a number of new methodological improvements for calculating structural and dynamical properties of proteins at long time-scales. First of all, we have developed a new mathematical approach to a classic geometrical problem in protein simulations, and demonstrated its superiority compared to existing approaches. Secondly, we have constructed a more accurate implicit model of the aqueous environment, which is of fundamental importance in protein chemistry. This model is computationally much faster than models where water molecules are represented explicitly. Finally, in collaboration with the group of structural bioinformatics at the Department of Biology (KU), we have applied these techniques in the context of modeling of protein structure and flexibility from low-resolution data.

Dansk resumé

Proteiner spiller en fundamental rolle i stort set alle processer i levende organismer. Nogle proteiner fungerer for eksempel som enzymer, der katalyserer en bred vifte af livsvigtige reaktioner, andre medierer vekselvirkningen mellem cellen og dens omgivende miljø eller optræder som regulerende molekyler i de cellulære processer. Nylige studier har påvist at proteiners specifikke biologiske funktion og aktivitet er tæt knyttet til deres strukturelle og dynamiske egenskaber. I princippet kan disse egenskaber studeres ved hjælp af computermodeller. Desværre optræder mange af de biologisk relevante strukturelle overgange i proteiner på tidskalaer som ikke er tilgængelige med nuværende simuleringsteknikker, da disse er beregningsmæssige omkostningsfulde. I denne afhandling præsenterer jer en række nye simuleringsteknikker til at beregne proteiners strukturelle og dynamiske egenskaber

på lange tidsskalaer. Først og fremmest har vi udviklet en ny matematisk løsning til et klassisk geometrisk problem i proteinsimuleringer, som udgør en væsentlig forbedring i forhold til eksisterende tilgange. Dernæst har vi konstrueret en mere præcis implicit model for det omgivende vands vekselvirkning med proteinet. Denne model er beregningsmæssig langt hurtigere end modeller, hvor vandmolekylerne er repræsenteret eksplicit. Endelig har vi i samarbejde med gruppen for strukturel bioinformatik på KU anvendt disse teknikker til at modellere proteiners struktur og fleksibilitet ud fra lav-opløsnings data. Samlet set bidrager mit arbejde således til bedre at beregne proteiners dynamiske egenskaber, hvilket har potentielle anvendelsesmuligheder indenfor en lang række bio-relaterede felter.

Contents

1

Introduction

Since their introduction in the 1950s [1, 2], theoretical methods and computational techniques have provided powerful insights into the nature of biomolecular systems. Despite the shortcomings and the limitations of current methodologies, molecular simulations can address questions of great biological relevance, ranging from elucidating the mechanism of protein function at atomic detail to studying the protein folding process [3, 4]. In this respect, atomistic molecular mechanics simulation is the most promising method for giving a useful and accurate description of structure and function of proteins. There are a number of challenges connected to such simulations. One is the design of an accurate mathematical and physical description of the interactions between atoms. Secondly, the accessible conformational space of biomolecules is so vast that simulations of large-scale, long-time configurational changes require a substantial computational effort. Despite of impressive hardware and software advances, computer power alone is often not sufficient for simulations to access biologically-relevant time scales. In the attempt to circumvent this problem, a large number of enhanced sampling techniques and coarse-grained models are being developed.

In the present thesis I describe the design and use of methodologies for conformational sampling of proteins. The development and assessment of a Monte Carlo (MC) method suitable for atomistic simulations of macromolecules (CRISP) constitutes the core of the PhD project (section 2.1). CRISP is one of the building blocks of PHAISTOS, a general-purpose software package for conducting MC simulations of proteins (section 2.2). The method is applied in the context of protein structure determination (section 2.3), to characterize the flexibility of globular proteins (section 2.4) and

served as a tool for elucidating the use of knowledge-based potential in MC simulations (section 4).

Monte Carlo simulations of biomolecules are typically performed using a simplified representation of the aqueous environment, in which an effective energy mimics the average influence of the solvent. The design of an accurate and computationally efficient representation of the water effects is a non-trivial theoretical problem. In the attempt to improve the current methodologies, I have optimized the parameters in a popular implicit solvent model using fully atomistic, state-of-the-art molecular dynamics simulations (section 2.5). To reach this goal, I have used a recently described coarse-graining technique based upon the minimization of an entropy-related function.

The following sections serve as an introduction to the research articles presented in Chapter 2. A special emphasis is given to the topics of the studies in sections 2.1 and 2.5, as they represent the two central works of this dissertation. A short introduction to protein structure, function and dynamics is given in the first part. In section 1.2, I give an overview of MC methods for simulations of biomolecules. The problems connected with the MC approach are presented, together with the methodologies proposed in literature to overcome these difficulties. Finally, the implicit solvent model and the coarse-graining technique used in the work of section 2.5 are introduced in the rest of the chapter (sections 1.3 and 1.4).

1.1 Protein structure, function and dynamics

Proteins play a fundamental role in virtually every process within living organisms. For example, some proteins act as enzymes, catalyzing a wide range of reactions necessary for life, others mediate the cell interaction with the surrounding environment and still others have structural or mechanical functions.

A protein is a linear polymer chain of amino acid assembled together using information encoded in genes. During and after protein synthesis, polypeptide chains often fold to assume their stable, biologically active functional form, called the *native state*. This process, known as *protein folding*, is probably the most significant example of conformational rearrangement in proteins.

The native state itself is not a static conformation, but a highly dynamic entity [5]. As a growing number of studies suggest, the inherent flexibility of a protein is intimately

related to its function, and holds the key to a number of biochemical processes such as signal transduction, antigen recognition, protein transport and enzyme catalysis [6]. From an experimental point of view, it is not yet possible to directly follow protein dynamics at atomic detail. Instead, it is possible to measure physical properties of the system from which the dynamics can be inferred. Computer simulations have the great advantage over real experiments as they can provide an atomic-resolution structural description of the conformational states, together with their relative probabilities and the energy barriers between them. Obtaining an atomic-detailed, complete *in silico* characterization of protein dynamics is, however, a highly non-trivial problem. In first place because it requires an accurate description of all the interactions between the particles in the system (*i.e.* a force-field). Secondly, because the typical time-scales over which many biological processes of interest occur are often not amenable by standard computational techniques (Fig. 1.1). In the light of these considerations, the effort of the scientific community is devoted to the improvement of both accuracy and precision (*i.e.* sampling efficiency) of molecular simulations (Fig. 1.1).

1.2 Monte Carlo simulations of proteins

The aim of a Monte Carlo simulation is the calculation of equilibrium properties of a system of interest. In the context of molecular simulations of proteins, the Monte Carlo method allows to sample conformations of the polypeptide chain according to a statistical ensemble distribution and to calculate expectation values as averages over sampled configurations.

Monte Carlo is not the most common choice for simulating biomolecules, as molecular dynamics (MD) is assumed to be superior for systems that are characterized by large correlations between many degrees of freedom [12]. One of the main difficulties of existing MC techniques is their inefficiency in dense environment, such as the native state of globular proteins. Here, even small variations in the degrees of freedom often result in steric clashes, and therefore in the rejection of the trial configurations (Fig. 1.2). Similarly, when using explicit solvent representation, any large-scale move altering the solute coordinates without also moving the solvent particles is likely to result in a substantial overlap of atoms.

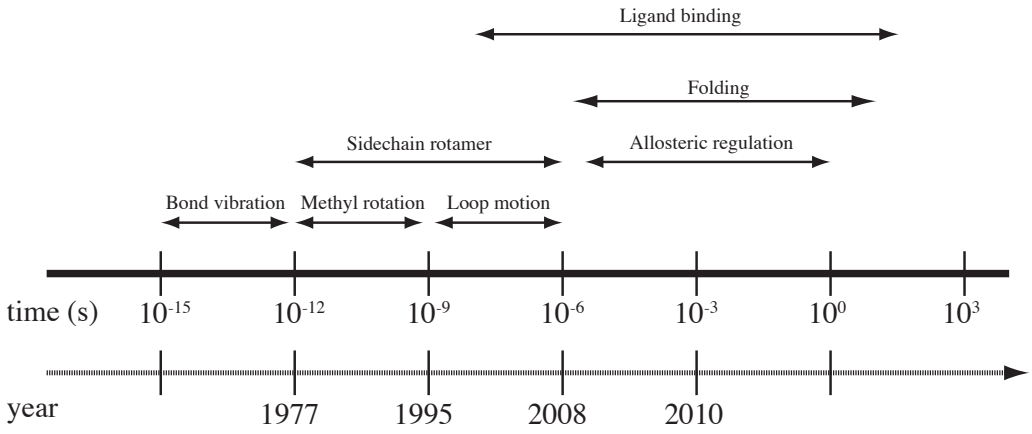


Figure 1.1: - Schematic illustration of the typical time-scales associated with conformational rearrangements in proteins. Bond vibrations and librations represent the fastest motions, occurring on the fs to ps time-scale. Side-chain rearrangements (methyl group rotations or rotameric transitions) and local backbone movements such as loop motions, take place on a longer time span, ranging from ps up to μs. The structural transitions directly connected to the different biological function of proteins (*e.g.* ligand binding or allosteric regulation) range from small motions of few amino acids in the binding site to domain re-orientation [7], and occur over a wide temporal window. Lastly, protein folding often involves large rearrangements of the whole polypeptide chain, and the very fastest known protein folding reactions are complete within a few microseconds [8], while time-scales of milliseconds are the norm. The bottom axis shows the evolution of the time-scales accessible to atomistic molecular dynamics simulations through the last four decades. 1977: $\approx$ 10 ps MD simulation on the bovine pancreatic trypsin inhibitor (BPTI) protein [9]. 1995: 5.3 ns MD simulations on chymotrypsin inhibitor 2 (CI2) [10]. 2008: 10 μs simulation on a WW domain mutant. 2010: 1ms MD simulation on BPTI [11].

Figure 1.2: Steric clash - A rotation of a single dihedral backbone angle by 2° at residue 36 on the native state of protein G (pdb code 1GB1), produces the structure colored in orange. The energy difference between the two structures calculated using the $OPLS_{AA}$ [13] potential is $\Delta E > 15000$ kcal/mol. This large energy difference is mainly due to electrostatic interactions between side-chains (not shown in this representation).

Another reason for the infrequent use of MC for protein simulations is the scarcity of specifically-tailored software [14, 15, 16], compared to the large number of widely used and established packages for conducting molecular dynamics simulations, such as GROMACS [17], NAMD [18], AMBER [19] and CHARMM [20].

Being free from following the dynamics of the system, however, Monte Carlo is a powerful tool for accessing large-scale, long time-scale conformational transitions. Moreover, in its standard form, MC does not require calculation of derivatives of the potential energy function. Therefore, it is easier to introduce holonomic constraints, non-differentiable potentials, generalized ensembles and non-physical weighting for enhanced sampling, although corresponding molecular dynamics methods have been introduced [21, 22, 23, 24]. Important biological processes that have proven difficult to treat using atomistic MD simulations, have been studied by means of MC methods. Notable examples are the aggregation of peptides [25, 26], the dimer formation of intrinsically disordered proteins [27] and the membrane absorption and insertion of peptides [28, 29].

Whether to use MD or MC for conformational sampling of proteins is problem-specific. Although a general rule does not exist, MC has an edge for simulations of short peptides and with implicit water representation [30], while it is seldom used for conformational sampling of native globular proteins [31, 32, 33, 34].

1.2.1 Biased Monte Carlo

In the standard Markov chain Metropolis scheme [1], the system is randomly perturbed and the trial move from the current microstate c to a new state n is accepted with a probability given by

$$a(c \to n) = \min\left\{1, \frac{p(n)}{p(c)}\right\} \tag{1.1}$$

where $p(x)$ is the probability of observing a given conformation x (*e.g.* the Boltzmann weight in the canonical ensemble). In the context of protein simulations, any large perturbation to the chain, such as a random rotation of a dihedral backbone angle, is likely to be rejected (Fig. 1.2). In order to enhance the sampling efficiency of MC on such systems, it is common practice to introduce non-random updates. To satisfy detailed balance, however, it is necessary to compensate for the bias introduced, by modifying the standard Metropolis acceptance rule

$$a(c \to n) = \min\left\{1, \frac{p(n)s(c \to n)}{p(c)s(n \to c)}\right\} \tag{1.2}$$

Here, $s(c \to n)$ is the probability of selecting the update $c \to n$ and $s(n \to c)$ is the probability for the reverse move $n \to c$. Typically, the efficiency of MC simulations can be enhanced by choosing a selection probability s that includes some prior knowledge of the target distribution p, for example by designing moves that alter only the *soft* degrees of freedom of the chain [35] or that take in account correlations or specific configurational propensities of the system [36].

A common prescription in Monte Carlo simulations is that the collection of moves (moveset) should be as diverse as possible, *i.e.* each type of move should involve different degrees of freedom and/or different scales of motion, compatibly with the system and the problem of interest. For this reason, a large variety of approaches has been proposed in literature, ranging from simple updates of a single dihedral backbone angle (Fig. 1.3 a), to local backbone rearrangements (Fig. 1.3 b-c), and to rotations of side-chain angles (Fig. 1.3 d).

An introduction to the MC algorithms specifically designed for off-lattice Monte Carlo simulations of biomolecules is given in the next part of this section. With the purpose of introducing the study presented in section 2.1, a broad overview of the different techniques for generating local Monte Carlo trial configurations is presented.

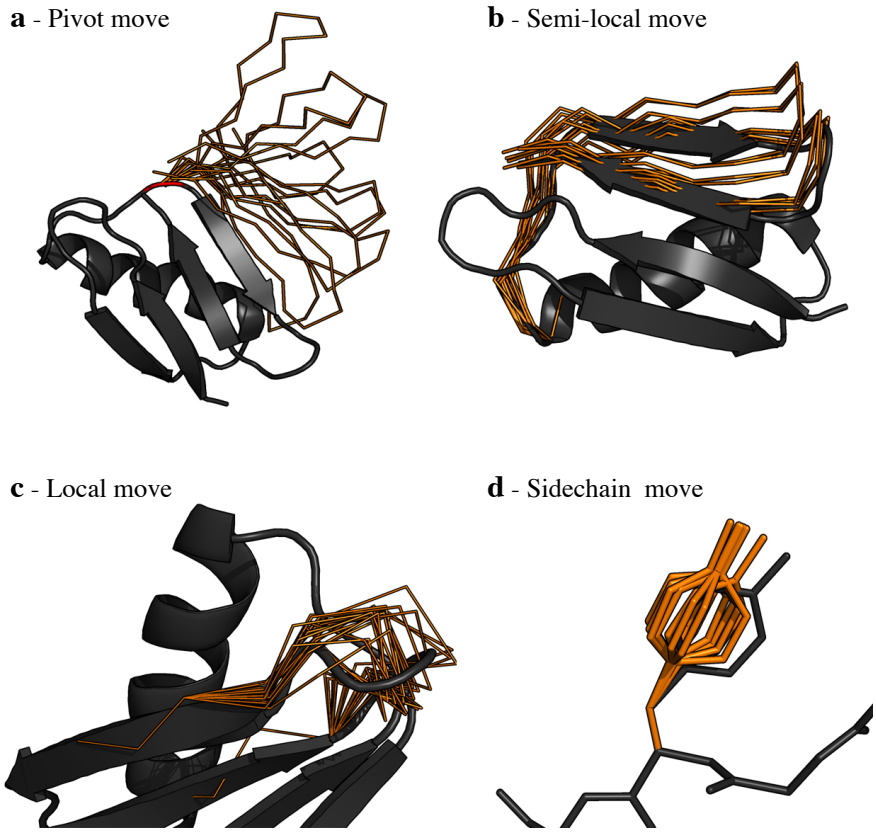


Figure 1.3: Monte Carlo moves - Illustrative representation of moves for Monte Carlo sampling of proteins. Trial updates to the original chain configuration (black) involve different degrees of freedom and different scales of motion, resulting in new configurations colored in orange. (**a**) Large structural rearrangements can be obtained by performing pivot moves, consisting in a rotation of a single dihedral backbone angle. (**b**) Semi-local moves also involve the rotation of dihedral backbone angles, but produce less drastic changes compared to pivot moves. Such updates can be obtained by introducing a bias that favors small displacement of the chain terminus, as in the biased Gaussian step move [37]. (**c**) Local moves alter only a restricted part of the protein backbone, leaving the rest of the chain unaffected. (**d**) The sampling of internal degrees of freedom in side-chains (*e.g.* χ angles) is performed by side-chain moves. For illustrative purposes, the different MC moves are here shown in separate panels and involve only specific residues. In a typical Monte Carlo simulation, however, these moves are combined and uniformly applied to the chain, so as to sample all degrees of freedom

Secondly, I introduce the concept of generative probabilistic models (GPMs) of local structure, *i.e.* probabilistic distributions that allow to sample trial configurations according to the specific angular propensities of proteins. Two possible applications of the approach are presented in sections 2.3 and 2.4. Finally, a number of alternative techniques are briefly discussed.

1.2.2 Local moves

A common approach in Monte Carlo sampling of flexible chain molecules is to introduce *local* moves, *i.e.* trial updates that alter only a restricted part of the chain, in order to reduce the probability of generating steric clashes and therefore enhancing the acceptance rate. In polymer science many MC methods are based on the reptation model [38] where a monomer unit is removed from one end and pasted at the other end. This method, however, cannot easily be applied to heterogeneous polymers with different side-chains like proteins. Local conformational changes involving only a small number of monomer units can be obtained by using kink jump or crankshaft motions [39, 40], as shown in Fig. 1.4.

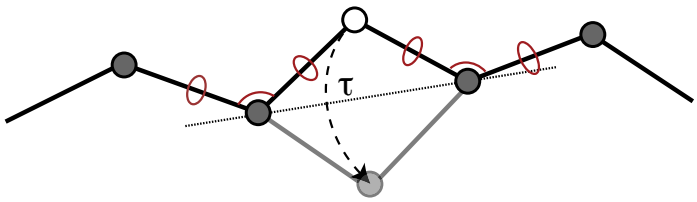


Figure 1.4: Crankshaft move - One backbone atom is chosen (white circle), and rotated by an angle τ around the axis connecting the adjacent atoms (dotted line) so as to vary the four dihedral and two bond angles shown in red.

Although crankshaft-type of moves have been employed for MC simulations of proteins, leading to successful applications in the context of protein modeling and for characterizing loop flexibility [41, 42], these moves involve the rotation of both bond and dihedral angles: this represents a problem for proteins, as backbone bond angles have little flexibility under physiological conditions [43].

These difficulties motivated the development of a large variety of more elaborate methodologies. Although all the approaches share many common features, local moves can be divided in two different classes: *concerted rotation* and *configurational-bias* methods.

The *concerted rotation* method consists in a cooperative change of several degrees of freedom that does not alter the conformation of the molecule outside a selected chain region. In the seminal paper by Gō and Scheraga [44] it was shown that the locality requirement (*i.e.* that the move is restricted to a selected region of the chain,

see Fig. 1.5) imposes 6 constraints among the n degrees of freedom involved in the move, thus giving $n - 6$ independent variables. In that study, the problem (also

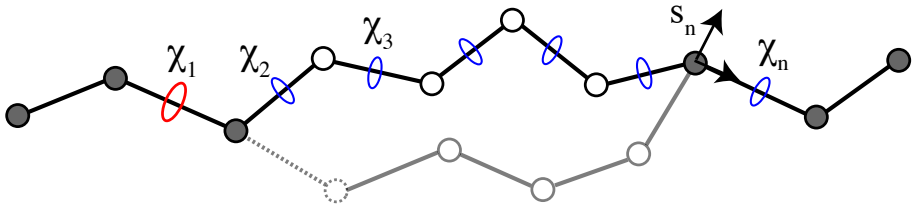


Figure 1.5: Loop closure - Illustration of the loop closure problem. The problem is to find new values for the n degrees of freedom $\chi_1 \ldots \chi_n$, which are compatible with the fixed bond length constraints, and that leaves the rest of the chain unaffected. A deformation cannot be confined to a local region for an arbitrary set of values for these n variables. Instead, some relations must exist among the n degrees of freedom. In order to meet the locality constraint, Gō and Scheraga [44] showed that it is sufficient to demand the location and orientation of the "last" local coordinate system (s_n) to remain fixed. The local coordinate system can be specified by six variables (three for the origin position and three for the Eulerian angles), thus introducing six constraints among the n degrees of freedom involved in the move. Therefore, the number of independent variables is $n - 6$. Based on this idea, Dodd *et al.* [45] devised the concerted rotation Monte Carlo move. The *driver* dihedral angle χ_1 (in red) is changed by a random amount, thus displacing the position of the adjacent atom (dashed circle). The values of the six remaining degrees of freedom (shown in blue) that restore the chain connectivity are determined by numerically solving a set equations for the six unknown variables, resulting in a new chain configuration (light gray).

known as the *loop closure* problem) was formulated as a set of equations for the six unknowns, reducible to a single equation of one variable. More recently, loop closure problems involving different monomeric units were solved exactly by reducing them to the determination of the real roots of a polynomial [46], or using results provided by the literature on inverse kinematics [47, 48]. In section 2.1 I demonstrate that a simple, analytical solution for loop closure is available, provided that bond angles are included as degrees of freedom (Fig. 1.6).

Based on the ideas of Gō and Scheraga, Dodd *et al.* [45] designed the concerted rotation Monte Carlo move, which consists in a rotation of seven adjacent dihedral angles (Fig. 1.5). Moreover, the authors showed that the solution to the loop closure problem entails a change in the variables used to describe the configuration space. Consequently, the Jacobian determinant of this transformation needs to be calculated and included

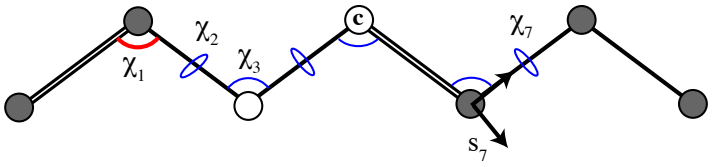


Figure 1.6: Loop closure in CRISP - Illustration of the loop closure problem in CRISP move, presented in section 2.1. Variations in the *driver* angles (for simplicity and without loss of generality, in this representation a single *driver* angle χ_1, shown in red, is used), produce the displacement of the adjacent atom on the right. Being the reference system s_7 fixed by the locality requirement, the loop closure simply consists in positioning the central atom (labeled **c**), resulting in new values for the six degrees of freedom shown in blue. As described in detail in section 2.1, this problem has an analytical solution based on simple geometrical considerations.

in the acceptance criterion in order to meet the detailed balance condition, an aspect that was neglected in previous studies [49]. The *concerted rotation* approach has been refined [50, 51, 52, 53], and modified in order to account for structural properties of polypeptides [54, 55] and nucleic acids [56, 57]. In 2003, Ulmschneider and Jorgensen [58] introduced a concerted rotation type of move (CRA) specifically designed for conformational sampling of proteins and showed their method to outperform the classic concerted rotation algorithm. The increased efficiency was obtained by including bond angle flexibility and by introducing a Gaussian bias favoring small angular variations. This idea, originally introduced by Favrin *et al.* [37] as a semi-local move, was used in the CRA method to ensure the existence of a solution for the loop closure equations (Fig. 1.7).

The *configurational bias* method consists of erasing an arbitrary section of a chain and regrowing it, segment by segment, until the original chain length of the molecule is restored (Fig. 1.8). While the original idea of regrowing a chain on a lattice from one random point to the end was proposed in the 50s by Rosenbluth and Rosenbluth [59], Siepmann and Frenkel introduced the configurational-bias Monte Carlo move [60], working out the necessary requirements that guarantee a Boltzmann-distributed sampling. The approach was then extended to continuum space [61, 62], and modified in order to produce strictly local perturbations [63]. Similarly to the case of concerted rotation techniques, a wide variety of improvements to the original method has been proposed. Specifically, Boltzmann weights and look-ahead strategies were employed for

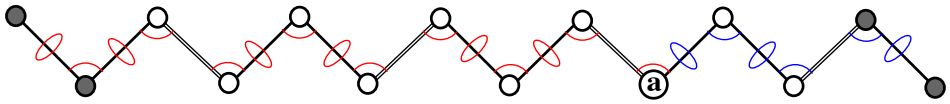


Figure 1.7: CRA move - The CRA move consists in a concerted rotation of bond and ϕ, ψ backbone angles in 5 consecutive residues, therefore involving the variation of $n-6=15$ *driver* angles (red lines). Subsequently, Gō and Scheraga's loop closure equations are used to determine the values for the 6 remaining degrees of freedom (shown in blue). If all driver angles are varied by a random amount, the position of the atom **a** is likely to change dramatically, therefore producing a geometrical configuration for which the chain closure is impossible under the constraint of fixed bond lengths (*i.e.* no solution for the loop closure is available). In CRA, this problem is circumvented by drawing driver angle variations from a Gaussian distribution that favors small displacements of atom **a**, as in the biased Gaussian step proposed by Favrin *et al.* [37].

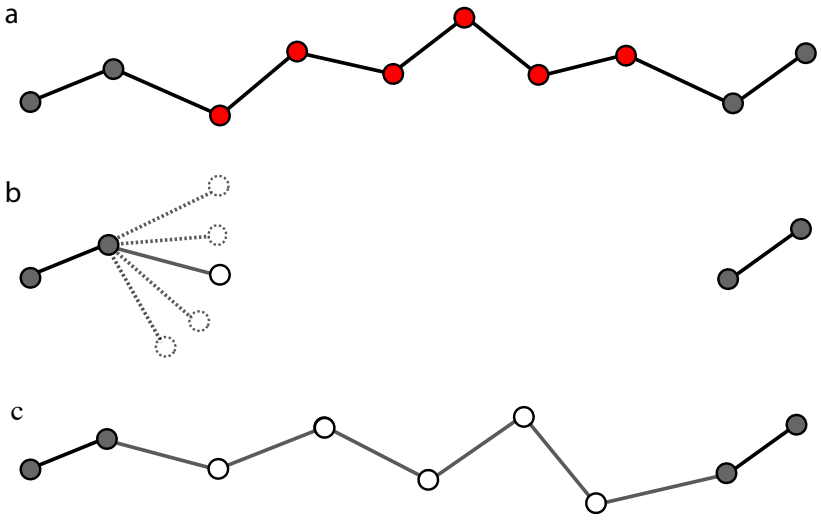


Figure 1.8: Configurational Bias local move - Pictorial representation of the re-bridging/internal configurational bias move. (**a**) An internal section of the chain (in red) is removed. (**b**) The chain is regrown site by site by selecting one new position from a set of trial configurations, until the chain connectivity is restored (**c**).

reducing the number of rejected configurations [64, 65, 66, 67, 68] or to encourage the regrowth towards the end of the segment [69, 70].

In summary, the design of a local move entails three distinct but connected problems:

1. devise a method to locally perturb the chain (either by using concerted rotations or regrowing techniques) under the constraints given by the molecule representation, such as fixed bond lengths, bond or dihedral angles.

2. design the move in such a way that the typical displacement of the atoms (step-size) can be controlled. It is worth noting that large updates are often rejected, and that exceedingly small changes lead to inefficient simulations. Having the possibility of tuning the step-size makes it possible to find the optimal trade-off in different simulation conditions (*e.g.* folded/unfolded states, all-atom/coarse grained representation, non-physiological temperatures).

3. determine the bias introduced by the move. Because of the non-random nature of the update, the selection probability is in general non-symmetric and, in order to preserve detailed balance, the ratio $\frac{p(n)}{p(c)}\frac{s(n\to c)}{s(c\to n)}$ in Eq. 1.2 has to be calculated.

Current methodologies only partially address these problems. For example, in classic concerted rotation methods controlling the stepsize is often problematic, as trial configurations typically introduce large dihedral angle ($> 40°$) [58] or bond angle ($> 5°$) variations with respect to the original structure. Moreover, none of those methods are known to work for dense systems such as the native state of globular proteins. In section 2.1 we present a novel concerted-rotation type of move (CRISP) specifically designed for atomistic simulations of proteins, that address many of the aforementioned problems. The algorithm is based on an analytical solution to the loop-closure problem (Fig. 1.6), making it possible to control the variations of all the degrees of freedom involved in the move. We show improved efficiency compared to the current state-of-the-art Monte Carlo techniques. Moreover, we prove CRISP to produce a near-native ensemble of a globular protein which is comparable to molecular dynamics simulations, both in terms of accuracy and efficiency. Considering that MC is expected to be highly efficient in non-compact states, this result strongly suggests that MC can be used as

valuable method in the study of macromolecules in atomic detail, offering a powerful alternative to molecular dynamics for probing large-scale, long time-scale conformational transitions.

1.2.3 Knowledge-biased moves

With the steadily growing number of experimentally solved protein structures, it has been possible to construct accurate models describing local conformational features of native proteins. Specifically, the well-known dihedral backbone [71, 72] and side-chain [73] angular propensities of proteins can be captured using for example fragment [74, 75] or rotamer libraries [76, 77]. These techniques usually rely on the clustering of conformations derived from experimental structures and are widely employed for protein structure prediction and modeling [78, 79, 80]. The accuracy of these methods has been improved by capturing the amino acid sequence and the secondary structure dependence of these libraries using hidden Markov models [81, 82]. Recently, several generative probabilistic models of local protein structure have been proposed [83, 84, 85, 86]. These models describe continuous probability distributions (*e.g.* of ϕ, ψ backbone angles [84] or χ side-chain angles [86]) that accurately capture the angular propensities of native proteins. More importantly, GPMs make it possible to sample trial configurations from such distributions, and to estimate the probability associated with any given conformation.

GPMs can be directly used in Monte Carlo simulations of proteins. As previously described, producing small perturbations is one possible way to enhance the efficiency of MC simulations. An even better candidate for $s(\cdot)$ in Eq. 1.2, however, is the one that approximates the target distribution $p(\cdot)$. In the case of chain molecules and at full atomic detail, the design of an easy-to-sample trial distribution $s(\cdot)$ is highly challenging, because of the high-dimensionality of the problem. In this context, GPMs can be directly used to propose tentative updates to the chain. Unlike the standard library-based approaches [82], in GPMs samples are drawn from a continuous probability distribution. Moreover, the probability associated with the proposed conformation is readily evaluated, and the bias introduced can be therefore compensated for, thus fulfilling the detailed balance condition. It should be also noted that the GPM Torus-DBN [84], that describes the (ϕ, ψ) backbone propensities, can be integrated into the

CRISP algorithm [87]. This combination makes it possible to propose trial configurations from the probability distribution in TorusDBN for a restricted region of the chain, thus combining the properties of GPM with the efficiency of CRISP.

By construction, GPMs mainly capture the local features of proteins. Non-local interactions can be described by specific models (*e.g.* hydrogen bond networks, NOEs distance restraints), that are easily incorporated as potentials in the MC framework. In this case, the acceptance criterion becomes

$$a(c \to n) = \min\left\{1, \frac{p(n)s_{loc}^{GPM}(c \to n)}{p(c)s_{loc}^{GPM}(n \to c)}\right\} \tag{1.3}$$

Here, s_{loc}^{GPM} is the GPM probability distribution from which trial configurations are drawn, while p is the target distribution.

Two applications of this approach are presented in Chapter 2. In section 2.3 TorusDBN [84] and Basilisk [86] (describing side-chain angle propensities) are used in combination with a network of user-defined, non-local restraints. The resulting method, TYPHON, is used to characterize the near-native state of globular proteins, and is proven to provide a simple, yet accurate method to investigate plausible conformational fluctuations in proteins. A similar approach is used in section 2.4, where TorusDBN and Basilisk serve again as trial distribution, while Nuclear Overhauser Effect measurements describe non-local interactions. This combination is successfully applied for protein structure determination on two model systems, and the results show improved efficiency and accuracy compared to the current state-of-the-art methodology.

In principle, GPMs can be used to enhance the sampling efficiency of MC simulations of proteins with classic all-atom physical force-fields. However, the probability distributions in GPMs are typically constructed using native structures and turn out not to be ideal to model the protein behavior under different conditions (*e.g.* simulations of unfolded states or at non-physiological temperatures). Moreover, ergodicity problems arise if the support of the trial distribution is only a subset of the full target distribution. Similar considerations apply to library-based approaches that were introduced with the purpose of improving the efficiency of MC simulations [88]. For a more complete introduction to probabilistic models and their use in MC simulations, I refer to the recent book of Hamelryck *et al.* [89].

1.2.4 Other techniques

Along different lines, a multitude of alternative approaches for Monte Carlo sampling of macromolecules have been proposed. MC updates involving collective motions of atoms like cluster algorithms were introduced for lattice spin models [90, 91], and later generalized for continuous systems such as Lennard-Jones fluids [92]. However, none of those methods are known to work for heterogeneous systems with long-range interactions such as proteins [15]. Around the pioneering idea of Noguti and Gō [32], normal modes were used to generate torsion space moves in Monte Carlo simulations, with successful applications on peptides [93]. Another possible approach is hybrid Monte Carlo, where a molecular dynamics simulation is run for a fixed length of time, and the final configuration is accepted or rejected with the standard Metropolis criterion [94, 95, 96]. Although hybrid Monte Carlo probably represents the most straightforward and general way to introduce the characteristic correlations of the system in the move set, the approach requires gradient calculations, therefore eliminating one of the advantage of MC over MD.

As a final remark, it is important to stress that local Monte Carlo updates can considerably enhance the simulation efficiency, but do not solve all sampling problems. As an example, when a large enthalpic barrier separates two potential energy minima, a stepwise evolution of the system is often not sufficient for barrier-crossing. This problem, of an ergodic nature, can be addressed by using multicanonical techniques [97, 98, 99].

1.3 Implicit solvent models

Many experimental and computational studies have underscored the role of the solvent environment in the description of structure and function of biomolecules [100, 101, 102]. The presence of water molecules is important in protein folding [103], determines the secondary structure propensities of peptides [104] and plays a role in complex formation and molecular recognition [105, 106]. For these reasons, the development of accurate computational water models is an important and active field of research [107]. The most widely used approach for modeling the aqueous environment is the inclusion of explicit water molecules, using for example the popular TIP3P [108] or SPC/E [109] models, in which a water molecule consists of three sites, representing the oxygen and the two hydrogens atoms. An alternative approach is to approximate the influence of the solvent by a potential of mean force that depends only on the atomic coordinates of the solute. This approach, called implicit solvation, has the advantage of being less computational demanding compared to explicit water simulations. Moreover, the reduced solvent viscosity and the smoothing of the energy landscape have the net effect of enhancing considerably the conformational search. The increased efficiency, however, comes at the price of reduced accuracy. An intrinsic limit of implicit solvent is that short-range effects mediated by few water molecules cannot be captured [110, 111]. Although the drawbacks of implicit solvent models should not be downplayed, these methods provide a fast and approximate way to address a variety of thermodynamic problems related to solvation of macromolecules [112].

Statistical mechanics is the natural theoretical framework for deriving a formulation of the approach. Assuming the potential energy, U, of a system to be additive, one can perform the following decomposition:

$$U(X,Y) = U_{mm}(X) + U_{ms}(X,Y) + U_{ss}(Y) \tag{1.4}$$

Here, the three terms represent the intra-molecular U_{mm}, the solute-solvent U_{ms} and the solvent-solvent interactions U_{ss}, while X, Y are the solute and water coordinates, respectively. The potential energy uniquely dictates the probability $p(X,Y)$ of observing the microstate (X,Y) in a thermal bath at temperature T

$$p(X,Y) = \frac{1}{Z}\exp\left(-\beta U(X,Y)\right) \tag{1.5}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann's constant and Z is the partition function $Z = \int \exp(-\beta U(X,Y))dXdY$. The central idea of implicit solvent is to define an effective energy $W(X)$ that depends only on the atomic coordinates of the solute. Formally, this is done by "integrating out" the solvent degrees of freedom in Eq. 1.5:

$$p(X) = \frac{\exp(-\beta W(X))}{Z_W} \equiv \int p(X,Y)dY = \frac{1}{Z}\int \exp(-\beta U(X,Y))dY \tag{1.6}$$

where Z_W is a suitable normalization constant. Using Eq. 1.4 and multiplying/deviding by $Z_s = \int \exp(-\beta U_{ss}(Y))dY$ we write

$$\begin{aligned} p(X) &= \frac{1}{Z}\exp(-\beta U_{mm}(X))\int \exp(-\beta U_{ms}(X,Y))\exp(-\beta U_{ss}(Y))dY \\ &= \frac{Z_s}{Z}\exp(-\beta U_{mm}(X))\int \exp(-\beta U_{ms}(X,Y))\frac{\exp(-\beta U_{ss}(Y))}{Z_s}dY \\ &= \frac{Z_s}{Z}\exp(-\beta U_{mm}(X))\langle \exp(-\beta U_{ms}(X))\rangle_s \end{aligned} \tag{1.7}$$

with the average being taken over solvent configurations [113]. This ensemble average is related to the free energy of solvation ΔG^{solv} [114] as

$$-\beta \Delta G^{solv} = \ln \langle \exp(-\beta U_{ms})\rangle_s \tag{1.8}$$

Therefore, the potential of mean force W can be written as

$$W(X) = U_{mm}(X) + \Delta G^{solv}(X) \tag{1.9}$$

Many of the classical mechanics force-fields describe the intra-molecular interactions U_{mm}, while the aim of implicit solvent is to provide an approximation for the free energy cost of solvating the molecule ΔG^{solv}.

The development of implicit solvent models for simulations of biomolecules has progressed mainly along three different lines of research: surface-area approaches [115, 116], solvent-exclusion [113, 117] and continuum electrostatics models [118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130]. The first two approaches assume the solvation free energy to be modeled by some physically sensible functional form, and the parameters of the model are adjusted such that specific experimental data are correctly reproduced. Although computationally very efficient, these methods lack a rigorous treatment of long-range electrostatic interactions. This important aspect

is more accurately described by continuum electrostatics models, where the solute is represented as a low-dielectric cavity embedded in a high-dielectric solvent. The corresponding (electrostatic) solvation free energy is calculated either by solving the Poisson-Boltzmann equation [118], or estimated using Generalized Born (GB) methods [119, 123]. Because of its relatively good accuracy, the latter method is recognized as a prime choice for the implicit treatment of solvent in biomolecular simulations [112], and a large variety of GB-based approaches can be found in literature [120, 121, 122, 124, 125, 126, 127, 128, 129, 130]. It should be noted, however, that the computational efficiency of standard GB methods is in many cases comparable to explicit water simulations [107].

In section 2.5 we propose a new parameterization for the Gaussian-exclusion implicit solvent model EEF1 [113], which we briefly introduce in the next part of this section.

1.3.1 EEF1: an effective energy function for proteins in water

In the EEF1 effective energy function the solvation free energy of the solute is modeled as a sum over atomic contributions

$$\Delta G^{solv} = \sum_i \Delta G_i^{solv} \tag{1.10}$$

where the individual terms, ΔG_i^{solv}, are given by the solvation free energy ΔG_i^{ref} minus the reduction due to the presence of surrounding groups

$$\Delta G_i^{solv} = \Delta G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j \tag{1.11}$$

ΔG_i^{ref} were obtained by dissecting the experimental free energy for a set of model compounds into group contributions [131] and the sum is performed over the groups j with volumes V_j around i. Finally, the solvation free energy density $f_i(r_{ij})$ is assumed to be a Gaussian function of the distance r_{ij}

$$f_i(r) 4\pi r^2 = \frac{2}{\sqrt{\pi}} \frac{\Delta G_i^{free}}{\lambda_i} \exp\left\{ -\frac{(r - R_i)^2}{\lambda_i^2} \right\} \tag{1.12}$$

this functional form is such that the volume integral over the first solvation shell of thickness λ (*i.e.* from $r_i = R_i$ to $r_i = R_i + \lambda_i$, where R_i is the van der Waals radius)

accounts for the 84% of the solvation energy. ΔG_i^{free} corresponds to the free energy of solvation of the isolated atom. In EEF1 this value is determined by requiring the solvation energy of deeply buried groups in the protein CI2 to be zero. λ was taken to be the thickness of one hydration shell (3.5 Å) except for ionic groups, for which a value of 6Å was used. In EEF1, a modified version of the united-atom CHARMM 19 energy function [132] describes the solute-solute interactions. Ionic charges are neutralized, and a distance-dependent dielectric constant is used to approximate electrostatics effects. Despite of the crudeness of the latter assumption, the EEF1 effective energy function is still widely and successfully used in a number of different contexts, such as folding simulations [133], protein structure prediction [134] and models solvent effects in the popular Rosetta software [135].

1.4 Coarse-graining Techniques

As already pointed out, atomistic molecular simulations with explicit solvent are recognized as the method of choice for providing an accurate description of protein structure, dynamics and function. The high computational cost of the approach, however, often poses a practical limit to its applicability. In parallel to the development of enhanced sampling techniques, several so-called *coarse-grained* (CG) approaches have been introduced. The goal of CG methods is to devise a simpler description of the effective interactions between particles, while retaining the ability of the model to reproduce the properties of the system. Typically, grouping atoms in fewer sites is a common way to reduce the problem complexity. Given a simplified representation, different methods can then be used to devise a proper description of the interactions between the CG structural units.

One possible coarse-graining method is given by the reverse Monte Carlo approach [136], where iterative Monte Carlo-based adjustments to the CG potential parameters are made, so as to correctly reproduce some quantity of interest (*e.g.* pair distribution function consistent with experimental measurements). A different strategy, devised by Ercolessi and Adams [137], constructs a CG potential by applying a force-matching procedure. The approach is based on the idea of minimizing the mean squared difference χ^2 between the forces $\mathbf{F}^0$ exerted on CG sites observed in an all-atom reference simulation, and those determined by the CG force $\mathbf{F}^{CG}$ at the same CG configuration

$$\chi^2 = \langle \frac{1}{3N} \sum_{i=1}^{N} |\mathbf{F}_i^0 - \mathbf{F}_i^{CG}(\bar{\eta})|^2 \rangle \tag{1.13}$$

Here, the summation runs over the N coarse-grained sites and the average is taken over all the atomic configurations used in the fit, while the optimization is carried out adjusting the set of parameters $\bar{\eta}$ in the CG force-field. The approach has been refined and successfully applied for developing CG models of liquid water [138], lipid bilayer [139] and peptides [140].

Recently, Shell [141] developed a multiscale coarse-graining approach that relies upon the minimization of an entropy-related objective function called relative entropy (RE). In the study presented in section 2.5, we extend the effective energy function EEF1, which was originally based on the united-atom CHARMM 19 force-field [132],

to the all-atom description of CHARMM 36[142]. The significant differences in the molecular representation as well as in the parameterization of the two force-fields does not allow a direct transfer of the EEF1 model parameters. Therefore, we devised a modified version of EEF1 were the model parameters are adjusted using the relative RE approach, which is presented in the remainder of this section.

1.4.1 Relative entropy approach

The RE approach is based on the minimization of the relative entropy S_{rel} (also known as Kullback-Liebler divergence) between the ensemble generated in an all-atom (AA) and in a coarse-grained (CG) simulation. From an information-theory perspective, the relative entropy quantifies the overlap between the two configurational ensembles AA and CG, which is linked to the amount of information lost due to coarse-graining. More precisely, the relative entropy S_{rel} is defined as

$$S_{rel} = \sum_R p_{AA}(R) \ln \frac{p_{AA}(R)}{p_{CG}(R)} \tag{1.14}$$

where $p(R)$ is the probability of a particular configuration R, and the sum proceeds over the AA microstates. By definition, the coarse-grained model has less degrees of freedom than the all-atom one, meaning that p_{CG} is a function of r only, where $r = r(R)$ is some given dimension-reducing mapping from the AA coordinates R onto the CG coordinates r [1]. This implies that

$$p_{CG}(R) = \frac{p_{CG}(r(R))}{\Omega(r(R))} \tag{1.15}$$

Here $\Omega(\tilde{r}) = \sum_R \delta_{\tilde{r},r(R)}$ is the degeneracy for any given CG state $\tilde{r}$, where the summation is performed over the AA configurations and δ is the Kronecker delta. Therefore, the relative entropy S_{rel} can be expressed as

$$S_{rel} = \sum_R p_{AA}(R) \ln \frac{p_{AA}(R)}{p_{CG}(r(R))} + \sum_R p_{AA}(R) \ln \Omega(r(R))$$

[1] In the study of section 2.5, the AA ensemble is obtained from explicit water simulations, while the CG ensemble is given by the solvent-exclusion model previously described. In this case, $R = (X, Y)$ where X and Y are the solute and solvent coordinates, respectively, while the mapping function is simply given by $r = r(X, Y) = X$.

$$= \sum_R p_{AA}(R) \ln \frac{p_{AA}(R)}{p_{CG}(r(R))} + \langle S_{map} \rangle \tag{1.16}$$

It is worth highlighting that the average entropy $\langle S_{map} \rangle$ does not depend on the CG ensemble beyond the specification of the mapping function.

In the canonical ensemble, the probabilities in Eq. 1.16 are linked to the potential energy of the CG an the AA model (U_{CG} and U_{AA} respectively) as in Eq. 1.5, and the relative entropy can be expressed as

$$S_{rel} = \beta(\langle U_{CG} - U_{AA} \rangle_{AA}) - \beta(A_{CG} - A_{AA}) + \langle S_{map} \rangle \tag{1.17}$$

where $A = -k_B T \ln Z$ is the free energy and $\beta = (k_B T)^{-1}$ is the inverse temperature T multiplied by the Boltzmann's constant, k_B. In this formulation, calculating S_{rel} requires the impractical estimation of free energies. Assuming the CG potential to be function of some parameters η, however, the derivatives of the relative entropy with respect to η can be expressed as simple averages over the CG and AA ensembles

$$\begin{aligned} \frac{\partial S_{rel}}{\partial \eta} &= \beta \langle \frac{\partial U_{CG}}{\partial \eta} \rangle_{AA} - \beta \langle \frac{\partial U_{CG}}{\partial \eta} \rangle_{CG} \\ \frac{\partial^2 S_{rel}}{\partial \eta^2} &= \beta \langle \frac{\partial^2 U_{CG}}{\partial \eta^2} \rangle_{AA} - \beta \langle \frac{\partial^2 U_{CG}}{\partial \eta^2} \rangle_{CG} + \beta^2 \langle \frac{\partial U_{CG}}{\partial \eta}^2 \rangle_{CG} - \beta^2 \langle \frac{\partial U_{CG}}{\partial \eta} \rangle_{CG}^2 \end{aligned}$$

Hence, standard numerical techniques can be employed to minimize the relative entropy with respect to the model parameter, for example by iterative application of the Newton-Raphson update rule

$$\eta_{k+1} = \eta_k - \gamma \left[\frac{\partial^2 S_{rel}}{\partial \eta^2} \right]^{-1} \left[\frac{\partial S_{rel}}{\partial \eta} \right] \tag{1.18}$$

From a practical point of view, performing the minimization (especially in a multi-dimensional parameter space) can be problematic. A detailed discussion of the numerical issues connected with the parameter optimization is presented in section 2.5, as well as in other studies by Shell and co-workers [143, 144, 145].

2

Research Articles

2.1 Subtle Monte Carlo Updates in Dense Molecular Systems

Density is commonly regarded as the limiting factor in making Monte Carlo (MC) useful for simulating biomolecules. While standard random updates are too inefficient for large systems and at high density, a clever choice of the update rule can considerably enhance the sampling efficiency of MC methods. In this research article we describe the design of a Monte Carlo move for producing local trial updates in a chain molecule. The approach is based on an analytical solution for the loop-closure problem, making it possible to control the variations of all the involved degrees of freedom. We demonstrate that CRISP reaches a comparable level of accuracy and efficiency as molecular dynamics simulations, and outperforms the current state-of-the-art MC methodologies. The supplementary material included after the manuscript contains a detailed description of the geometrical issues connected to the loop-closure problem and the complete derivation of the CRISP move. Moreover, detailed balance is proved and a number of geometrical and energetic aspects are discussed.
This is a joint first-author paper, and I was involved in every aspect of the work.

JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Subtle Monte Carlo Updates in Dense Molecular Systems

Sandro Bottaro,[*,†,∥] Wouter Boomsma,[*,†,‡,∥] Kristoffer E. Johansson,[§] Christian Andreetta,[§] Thomas Hamelryck,[§] and Jesper Ferkinghoff-Borg[*,†]

[†]Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

[‡]Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden

[§]Department of Biology, University of Copenhagen, Copenhagen, Denmark

Supporting Information

ABSTRACT: Although Markov chain Monte Carlo (MC) simulation is a potentially powerful approach for exploring conformational space, it has been unable to compete with molecular dynamics (MD) in the analysis of high density structural states, such as the native state of globular proteins. Here, we introduce a kinetic algorithm, CRISP, that greatly enhances the sampling efficiency in all-atom MC simulations of dense systems. The algorithm is based on an exact analytical solution to the classic chain-closure problem, making it possible to express the interdependencies among degrees of freedom in the molecule as correlations in a multivariate Gaussian distribution. We demonstrate that our method reproduces structural variation in proteins with greater efficiency than current state-of-the-art Monte Carlo methods and has real-time simulation performance on par with molecular dynamics simulations. The presented results suggest our method as a valuable tool in the study of molecules in atomic detail, offering a potential alternative to molecular dynamics for probing long time-scale conformational transitions.

1. INTRODUCTION

The conformational flexibility of molecules plays a central role in many important biological processes, including signaling, catalysis, regulation, and aggregation.[1−4] Structural and dynamical information of the conformational changes associated with these processes can be partly extracted from spectroscopic techniques, such as X-ray diffraction or nuclear magnetic resonance experiments (NMR).[4] Molecular simulations serve as an ideal complement to these techniques, by allowing the conformational variation to be studied at a detailed atomic level.

Molecular simulations, however, are faced with two main challenges: the design of an accurate energy function[5] and the construction of a sampling strategy capable of efficiently exploring the conformational space.[6] In the all-atom physical potentials usually employed in protein simulations, the energy landscape is rugged and complex due to the presence of a large number of protein−protein and protein−solvent interactions. For these systems, molecular dynamics (MD) is commonly considered the technique of choice.

The alternative approach to molecular simulation, Markov chain Monte Carlo (MC), has the potential to explore the energy landscape more rapidly than MD. In particular, the transitions between consecutive microstates in an MC simulation are not required to follow the dynamics of the system (i.e Newton's law). Using a scheme to accept/reject proposed updates to the chain, it is possible to generate conformations according to the Boltzmann distribution associated with the system. While the MC approach does not provide explicit real-time information, it allows for a rapid exploration of conformations separated by high-energy barriers (i.e., long time-scales) and thereby an efficient thermostatistical characterization of the system. This makes the Monte Carlo method extremely well suited for large scale simulations of, for instance, protein aggregation[7] or for exploring the conformational space of intrinsically disordered proteins,[8] both of which are still mostly intractable using MD. However, for the exploration of dense systems, where even small variations in the degrees of freedom (e.g., dihedral angles) are likely to introduce collisions in the molecule, MC simulations often perform poorly. In an attempt to alleviate this problem, many MC procedures extend their kinetics with so-called *local moves*, which produce subtle deformations in a small segment of the protein chain, while keeping the positions of all atoms outside the segment fixed.

The geometrical issues behind the local move problem were first studied by Gō and Scheraga in 1970.[9] On the basis of these considerations, Theodorou and co-workers developed the *concerted rotation* MC-move by working out the necessary requirements for detailed balance (the central condition to ensure Boltzmann distributed sampling).[10] The method works with seven adjacent dihedral angles along the chain. One of these angles is turned by a random amount, and the values of the six remaining angles are determined by numerically solving a set of equations, resulting in a new closed chain structure. Several variants of this original approach have been proposed.[11,12] The most recent is the CRA method,[13] in which increased efficiency was obtained by including bond angle variations[14] and imposing a locality constraint to raise the probability of chain closure, a technique originally introduced in the context of semilocal moves.[15]

Several alternative formulations of the local move problem have been proposed. The *configurational bias* method is based on the idea of regrowing a segment of a chain one atom at a time.[16,17] This approach has been extended with various look-ahead and

Received: September 14, 2011
Published: December 19, 2011

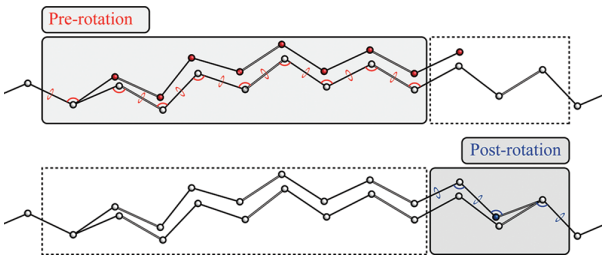


Figure 1. Illustration of the concerted rotation method. During prerotation, new values for the angles shown in red (light gray box) are proposed for a small segment of the chain, introducing a break of the chain. The role of the postrotation step (dark gray box) is then to find the necessary compensating changes in the six remaining degrees of freedom, labeled in blue, in order to return to a closed state of the chain.

biasing strategies to decrease the number of rejected growth attempts.[18,19] More recently, robotics-inspired methods have been employed to perform local backbone deformations[20] and to characterize the flexibility of protein loops.[21] As another alternative, an off-lattice version of the *crankshaft* move has been proposed.[22] The method consists of a rigid rotation of a chain segment around the axis defined by the C_α atoms delimiting the segment. An extension of this approach, the *backrub* move,[23,24] has led to successful applications in the context of both protein design and modeling.[25]

The crankshaft/backrub move stands out from the remaining methods for its simplicity and ease of implementation. However, the kinetics produced by this move are limited to hinge-like motions, which can potentially reduce the rate with which the move can decorrelate a structure. The remaining local move methods all introduce a break in chain-connectivity. This forces the methods to treat the placement of a subset of the atoms as a special case in order to maintain a closed chain, thereby introducing an asymmetry in the degrees of freedom involved in the move. Using Boltzmann factors or constrained proposals, it is possible to control the local geometry for the initial, stochastic part of the move. However, the final closure step will typically introduce unfavorable local structure in the chain, leading to an elevated rejection rate.

In the present study, we demonstrate that this problem constitutes one of the primary bottlenecks in current MC simulations of dense protein systems. We present a novel and efficient solution, in which the geometrical constraints are naturally incorporated in a proposal distribution. This leads to a concerted-rotation type Monte Carlo move, *CRISP* (Concerted Rotations Involving Self-consistent Proposals), which effectively proposes closed structures with user-controlled variations of all involved degrees of freedom. We demonstrate the correctness of the method and assess its efficiency by estimating the correlation time associated with the move. The results demonstrate that CRISP significantly outperforms the current state of the art MC methodologies. We proceed with a study of the native ensemble of ubiquitin. A comparison to X-ray and NMR experimental data shows that our improved sampling strategy enables us to cover the entire known conformational fluctuation spectrum of ubiquitin in solution, including several experimentally confirmed conformational switches. In addition to the clear performance improvement over existing MC methods, we demonstrate that our method has comparable real-time performance to MD on this system.

2. RESULTS

2.1. Method Overview. For simplicity, we will present the method in the context of protein molecules, although the basic principles apply to other chain molecules (Text S1 and S2, Supporting Information). A typical parametrization used in protein simulations is one with flexible dihedral angles and bond angles, but with bond lengths fixed. Given this parametrization, the local move problem can be phrased as follows: propose new values for all dihedral and bond angles in a region of a protein chain so that any atom position outside the region remains untouched. The requirement of chain integrity imposes a strong dependency among the degrees of freedom. From this perspective, the local move problem is essentially a matter of finding the cross-correlation between the degrees of freedom that fulfills the geometrical constraints given by the protein representation. In this paper, we will demonstrate how a probability distribution can be constructed that takes these dependencies into account. A natural framework for these considerations is that of the *concerted rotation*. [Note that it could also be formulated as a *configurational bias* move with the bias given by the derived probability distribution.] In the concerted rotation approach, a move is divided into a stochastic *prerotation* step followed by a deterministic *postrotation* step. During prerotation, new angles are proposed for a small segment of the chain, introducing a break of the chain. The postrotation step then closes the chain by finding the necessary compensating changes in the six postrotational degrees of freedom (Figure 1).

The derivation of our desired probability distribution is based on two observations. First, we note that given the described molecular chain representation, an exact, analytical solution for the postrotation problem can be derived (Text S1). This means that for any given value of the prerotated degrees of freedom, the resulting postrotation values can be determined with high efficiency and robustness. This solution represents a great advantage over other concerted-rotation methods by avoiding the tedious numerical resolution of a system of six equations in six unknowns.

The second step is the realization that the analytical solution allows us to express the coupling between pre- and postrotation as a linear transformation, which enables the construction of a probability distribution that controls both pre- and postrotational degrees of freedom as well as the necessary chain-closure constraints. To our knowledge, this is a novel mathematical description of the 40-year-old chain closure problem. Unlike previous approaches, it makes it possible, to first order, to directly sample closed chains. Since the complete derivation is quite involved, we only highlight the main features here and refer the reader to the Supporting Information for details (Text S2−S4).

To illustrate the nature of the procedure, we consider a local move where n degrees of freedom $\bar{\chi} = (\chi_1 \ldots \chi_n)$ of the chain backbone are modified, leading to a new conformation $\bar{\chi}'$. Angular variations $\delta\bar{\chi} = \bar{\chi}' - \bar{\chi}$ are drawn from a multivariate Gaussian distribution

$$p(\delta\bar{\chi}) \propto \exp\left(-\frac{1}{2}\delta\bar{\chi}^T \lambda \mathbf{C}_n \delta\bar{\chi} \right) \quad (1)$$

where the scalar parameter λ specifies the degree of locality, with increasing λ leading to smaller changes. $\mathbf{C}_n$ is an n-dimensional diagonal matrix introduced with the purpose of scaling, by a factor k, the allowed variations of bond and ω dihedral angles

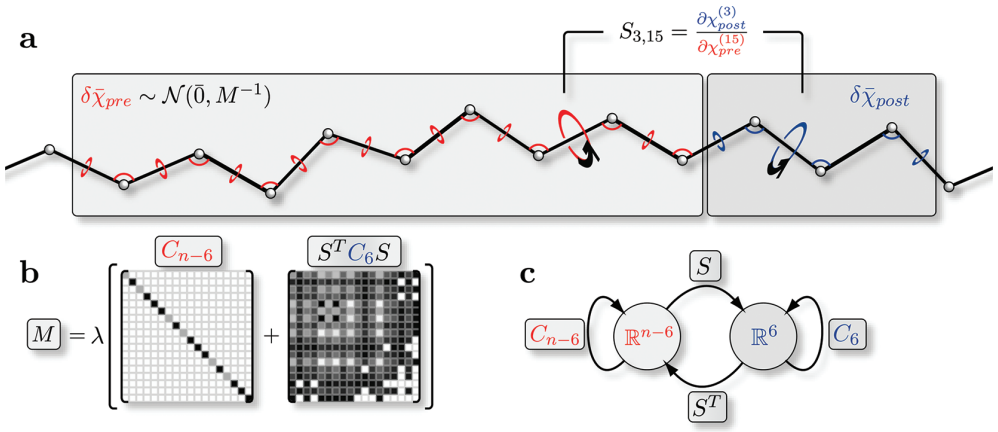


Figure 2. Graphical representation of the proposal probability distribution used in CRISP moves. (a) Angular variations are drawn from a normal distribution with mean $\bar{0}$ and covariance $\mathbf{M}^{-1}$. The $\mathbf{S}$ matrix couples the prerotational degrees of freedom (red) to the postrotational ones (blue). In the example shown in the figure, the matrix element $S_{3,15}$ reports the variation of the third postrotational angle upon a change in the prerotational degree of freedom at position 15. (b) $\mathbf{M}$ is a sum of a diagonal matrix $\mathbf{C}_{n-6}$ that controls the variations of the prerotational degrees of freedom, and $\mathbf{S}^T\mathbf{C}_6\mathbf{S}$. This last, nondiagonal matrix operates on $\delta\bar{\chi}_{\text{pre}}$ as shown in c: the $\mathbf{S}$ matrix first reports the changes $\delta\bar{\chi}_{\text{post}}$ upon the variation $\delta\bar{\chi}_{\text{pre}}$. The variations of postrotational angles, shown in blue, are then properly constrained via $\mathbf{C}_6$, and $\mathbf{S}^T$ maps back the postrotational changes to the prerotational, $n-6$-dimensional space. λ is a free parameter controlling the overall size of the move.

relative to φ,ψ angles:

$$C_{ii} = \begin{cases} k & \text{if } i \text{ is a bond or } \omega \text{ dihedral angle} \\ 1 & \text{if } i \text{ is a } \varphi \text{ or } \psi \text{ dihedral angle} \end{cases} \quad (2)$$

Due to the deterministic nature of the chain closure problem, the values of the six postrotational degrees of freedom $\bar{\chi}_{\text{post}} = (\chi^{(1)}_{\text{post}} \cdots \chi^{(6)}_{\text{post}})$ are determined by the remaining $n-6$ prerotational angles via our analytical solution. To first order, this allows us to express the variation of the six postrotational angles as a function of the prerotational variation, $\delta\bar{\chi}_{\text{post}} = \mathbf{S}\delta\bar{\chi}_{\text{pre}}$, where $\mathbf{S}$ is a $6 \times (n-6)$ matrix (Text S2). This information can be directly embedded in the proposal distribution of eq 1. As demonstrated in Text S3, the proposal distribution can now be written as

$$p(\delta\bar{\chi}_{\text{pre}}) \propto \exp\left(-\frac{1}{2}\delta\bar{\chi}^T_{\text{pre}}\lambda(\mathbf{C}_{n-6} + \mathbf{S}^T\mathbf{C}_6\mathbf{S})\delta\bar{\chi}_{\text{pre}}\right)$$
$$= \exp\left(-\frac{1}{2}\delta\bar{\chi}^T_{\text{pre}}\mathbf{M}\delta\bar{\chi}_{\text{pre}}\right) \quad (3)$$

Figure 2 illustrates the construction of the proposal function in eq 3. Angular variations for the prerotational degrees of freedom are drawn from a Gaussian distribution with mean $\bar{0}$ and covariance $\mathbf{M}^{-1}$ (Figure 2a). $\mathbf{M}$ is a sum of two terms (Figure 2b): the scaling diagonal matrix $\mathbf{C}_{n-6}$, acting on the prerotational angles, and $\mathbf{S}^T\mathbf{C}_6\mathbf{S}$. The latter, nondiagonal matrix carries the correlations between pre- and postrotational angles arising from the fixed bond-lengths constraint and from the restrictions given by the stereochemistry of the protein backbone. In other words, it operates on $\delta\bar{\chi}_{\text{pre}}$ as depicted in Figure 2c: The $\mathbf{S}$ matrix first reports the compensating changes $\delta\bar{\chi}_{\text{post}}$ upon the variation $\delta\bar{\chi}_{\text{pre}}$. The postrotational variations are then properly constrained via $\mathbf{C}_6$, and $\mathbf{S}^T$ finally maps the changes back to the $n-6$ dimensional space of the prerotation.

A proof of correctness of the first order approximation and a study of the range of its effectiveness is presented in Figures S1 and S2. We check the validity of the MC procedure, outlined in Text S4, by demonstrating detailed balance (Figure S3). The approach is further validated by demonstrating that, for long simulations on a small system, MC and MD methods produce comparable ensembles (Figure S4 and Table S1).

We proceed by establishing the performance of CRISP relative to two successful local move methods from the literature: the CRA concerted rotation method (CRA)[13] and a detailed balance version of the crankshaft-based *backrub* method (CRANKSHAFT)[24] (see Materials and Methods).

2.2. Controlled Variations. The main motivation for introducing local kinetics into a simulation is to increase sampling efficiency in dense systems, since nonlocal moves tend to propose a high number of self-colliding structures. However, while local moves successfully reduce the rate of self-collision, they are faced with a different problem: due to the strong chain-closure constraints, local moves will often introduce unfavorable values to a subset of the involved degrees of freedom, leading to an increase in the rate of rejection. We illustrate this problem using the CRA concerted rotation move. The method is constructed around the idea of limiting the movement of the end point of the prerotation (the breakpoint), in order to increase the probability of finding a solution for the postrotation. It is evident from Figure 3 that this strategy creates an imbalance in the move: constraining the displacement of the breakpoint is not sufficient to avoid significant fluctuations of the postrotational angles. In this case, the effect does not represent a significant problem for dihedral angles, but a typical change of 5° for *all* postrotational bond angles is dramatic, considering that the experimentally observed distribution width is ~2.7° for such degrees of freedom (Figure S5).[26] Other local move methods suffer from similar problems: in crankshaft-type moves, the bond angles surrounding the pivotal points will be subject to large fluctuations, while concerted rotation methods that do not include bond angles typically involve large jumps in dihedral angle values.[10,11] In contrast to existing methods, Figure 3 demonstrates that CRISP displays identical variations in pre- and postrotational degrees of freedom, *de facto* eliminating the asymmetry introduced by the

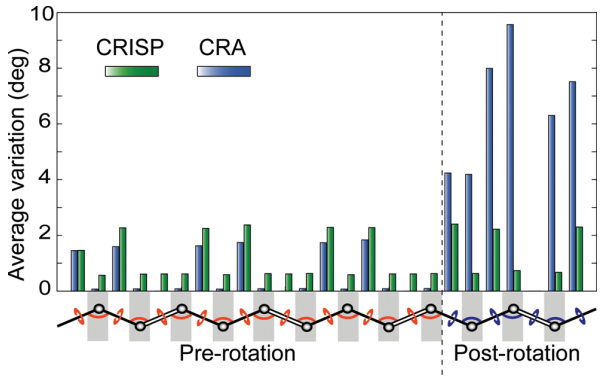


Figure 3. Average angular variations based on 5×10^4 attempted CRISP and CRA updates on ubiquitin. Each bar corresponds to the average variation of the degree of freedom shown in the chain below the histogram. In the prerotation (red angles), similar average angular variations are proposed by both methods. During postrotation (blue angles), large angular variations are introduced by CRA, due to the lack of a strategy controlling these degrees of freedom. Conversely, no imbalance between pre- and postrotation is observed when using CRISP.

chain-closure constraint. Note that the difference between bond angle and dihedral variations is user-defined (see eq 2).

2.3. Simulation Efficiency. The optimal way to quantify sampling efficiency is by measuring the correlation time associated with a given kinetic algorithm,[27] which represents the number of MC steps separating two independent samples. The correlation time allows us to compare the efficiency of the different methods and to establish the optimal values of the two free parameters of CRISP. Since obtaining converged estimates of the correlation time requires extensive simulations, we consider the equilibrium fluctuations around the stable helical state of the small peptide Ala_{14}.[13]

In Table 1, we report the correlation times in MC steps for the energy, τ_e, and the average correlation time, $\bar{\tau}_d$, of the 20 central dihedral angles when using CRISP, CRA, and CRANKSHAFT moves as described in the Materials and Methods. There is a difference in the dimension of configurational space explored by the three methods which should be taken into account when comparing the correlation times. Given the correlation time, the size of explored space can be estimated by calculating the standard deviation σ of the distribution for the energy and for dihedral angles (Table 1). CRISP shows a dramatic improvement of a factor of 15−20 in sampling efficiency compared to CRA as a consequence of the more appropriate treatment of the geometrical problem. Furthermore, the CRANKSHAFT move explores a conformational and energetic space which is ∼30% smaller than the one covered by CRISP for each degree of freedom, at a computational cost that is 2−3 times larger.

The two free parameters of CRISP, k and λ, were optimized with respect to the correlation time (Figure S6). In our experience, this setting is not sensitive to the type of protein being simulated, and all simulations in our study therefore use these values.

2.4. The Native Ensemble of Ubiquitin. While the Ala_{14} system is useful for the calculation of correlation times, it is not representative of the structural heterogeneity observed in native globular proteins. We therefore extend our analysis with a study of ubiquitin. Ubiquitin is a key to several cellular signaling networks[28,29] and is recognized by a broad variety of proteins

Table 1. Correlation Time τ in MC Steps and Standard Deviation σ of the Distribution for CRA, CRANKSHAFT, and CRISP Moves Calculated over 5 Independent Runs

	$\bar{\tau}_d$ (10^3 steps)	$\bar{\sigma}_d$ (deg)	τ_e (10^3 steps)	σ_e (kcal/mol)
CRA	13.2 ± 1.1	9.31	18.2 ± 3.3	3.04
CRANKSHAFT	1.9 ± 0.4	7.68	2.9 ± 0.9	2.42
CRISP	0.78 ± 0.01	10.67	0.85 ± 0.05	3.39

with high specificity. Furthermore, this protein is well characterized by NMR[30−33] and has been used extensively as a model system in previous computational approaches.[34−37] We use this as a model system for a comparison of CRISP to existing MC sampling algorithms, to MD, and to stochastic dynamics[38] (SD) simulations.

Structural fluctuations around the native state can be expressed as root mean squared fluctuations (RMSF), which measure the amplitude of movements of individual atoms around their equilibrium positions. The RMSF values generally grow with the simulation time, converging when the neighborhood of the local energy minima is exhaustively explored. We captured the global time evolution of this process by considering the sum of the individual RMSF values for all C_α atoms in the chain (cumulative RMSF). Although not as rigorous as the correlation time estimation, this procedure is useful to evaluate and compare the efficiency of different sampling techniques in the vicinity of the native state. In Figure 4, we show the simulation-time evolution of the cumulative C_α RMSF for the different methods. For each method, we report the average cumulative RMSF over 10 simulations performed at T = 300 K, starting from a relaxed state of the human ubiquitin X-ray structure (1UBQ). The MD/SD simulations covered 10 ns using the exact same force field and conditions (see Materials and Methods). For visualization purposes, the x axis for MD/SD is scaled to match the CPU time of the MC methods on the same machine.

Since detailed balance is fulfilled for all simulations (Figures S2 and S4), we expect the fluctuations of the different methods to eventually converge to the same levels. The observed differences thus reflect the degree of ergodicity obtained by the various sampling methods within the given simulation time. For CRA and CRANKSHAFT, the cumulative RMSF saturates at around 40 Å with a similar convergence time (Figure 4). The SD simulations are considerably faster but saturate approximately at the same level. Our CRISP method clearly outperforms the competing MC methodologies: the cumulative RMSF quickly crosses the 40 Å barrier, saturating at the same level as the MD simulations (∼50 Å). To further investigate the nature of these fluctuations, Figure 5 shows the converged RMSF profile per C_α atom for the different simulation methodologies. The fluctuations produced by CRISP are remarkably similar to MD, while the RMSF profiles of SD, CRA, and CRANKSHAFT are consistently lower.

As an experimental reference, we present the RMSF of two NMR-derived ensembles: MUMO (PDB code 2NR2[35]) and EROS (PDB code 2K39[33]), selected to represent the variation of experimentally based ensembles reported in the literature (Figure 6b). The fluctuations obtained with CRISP and MD are in good agreement with the experimental data. Specifically, the large variability observed in the $\beta_1-\beta_2$ loop, the C-terminal region of α_1, and the $\beta_3-\beta_4$ loop cover the main conformational variability observed in X-ray ubiquitin complexes, as represented

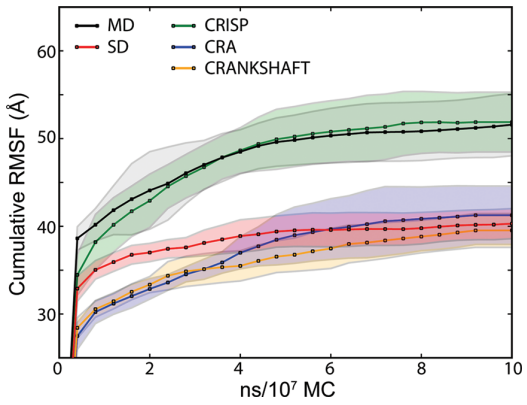


Figure 4. Time evolution of the cumulative C_α RMSF relative to MD simulations (in black/gray) compared with SD and with MC simulations using CRISP, CRA, and CRANKSHAFT as local moves around the native state of protein ubiquitin. The shaded regions show the standard deviation based on 10 independent runs.

by the zinc finger ubiquitin-binding domain of isopeptidase T (2G45) and the conjugating enzyme (E2) binding domain (1AAR)[33] (Figure 6c−d).

Large fluctuations in the MD and CRISP simulations are also observed in the α_1 N-cap region around GLU24 and in the $\beta_4-\alpha_2$ loop around GLY53. It is worth noticing that residual fluctuations in these regions are directly linked to a more subtle conformational switch consisting of the flipping of the ASP52/GLY53 amide plane (Figure 6e). This movement exposes the backbone CO of ASP52 to the exterior of the molecule, while the backbone NH of GLY53 forms an internal hydrogen bond with the side chain of GLU24, which changes rotamer state due to this interaction. This switch has recently been observed in a crystal structure of monomeric human ubiquitin,[39] and the flexibility of these residues was hypothesized to play a role when ubiquitin binds with deubiquinating enzymes. Notably, this backbone transition is not observed in the 10 ns SD or in the CRA/CRANKSHAFT simulations.

We stress that any such *in silico* observation is in principle a consequence of the applied force field, not the sampling method used. However, the fact that these transitions are not seen by all of the simulation methods within the same time frame using the same force field points to the importance of an efficient sampling strategy.

We analyze the different ensembles in detail by comparing the probability distributions of both dihedral backbone angles and C_α positions produced by the different methodologies. In Table 2, we report the Jensen-Shannon divergence (JSD) of the φ/ψ distributions between a reference ensemble $\overline{\text{MD}}$ and the individual trajectories (Figure S7). The $\overline{\text{MD}}$ ensemble is constructed using the samples from all conducted MD simulations. MD serves as a meaningful reference, as it represents the broadest and most complete representation of the near-native dynamics among the existing methods. The results show that CRISP and SD simulations reproduce the dihedral backbone distributions of $\overline{\text{MD}}$ with greater accuracy compared to CRA and CRANKSHAFT. The same scenario is observed when considering the Kullback−Leibler divergence (KLD) between C_α position distributions (Table 2), a recently proposed alternative measure of ensemble similarity.[40]

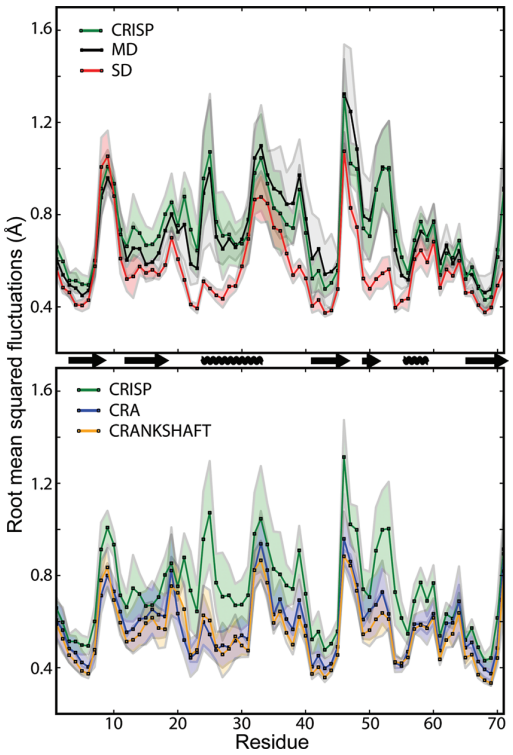


Figure 5. Ubiquitin RMSF from 10 ns MD simulations (in black/gray) compared to SD and to 9×10^7 long MC runs using CRISP, CRA, and CRANKSHAFT. The shaded regions show the standard deviation based on 10 independent runs.

3. DISCUSSION

The efficiency of Markov chain Monte Carlo protein simulations relies heavily on the kinetic algorithm used to probe the various possible conformational states of the molecule. In particular, densely packed systems typically require the presence of a move which restricts itself to modifying atom positions within a short stretch of the molecule. Designing such moves is a nontrivial task, due to the complex interdependencies among bond and dihedral angles that arise from constraining the end point of the modified stretch to be fixed. Our present study introduces a novel technique for incorporating these interdependencies directly into a proposal distribution. The resulting local move, CRISP, displays significant performance improvements compared to existing state-of-the-art MC methods.

It should be noted that efficient side-chain dynamics is an important prerequisite for obtaining the presented MC results. In particular, in the often dense hydrogen bond networks characterizing native proteins, one should ensure that an MC simulation includes side-chain moves that can break and form these bonds independently. We discuss this issue in greater detail in the Materials and Methods and in Figure S8. In addition, when conducting a local backbone move, the corresponding displacement of rigidly attached side-chain atoms can lead to self-collisions in the chain, and thus an elevated rejection rate. This consideration is of great importance especially for long side-chains at high densities. The presented move could therefore potentially be improved by including constraints from side-chain

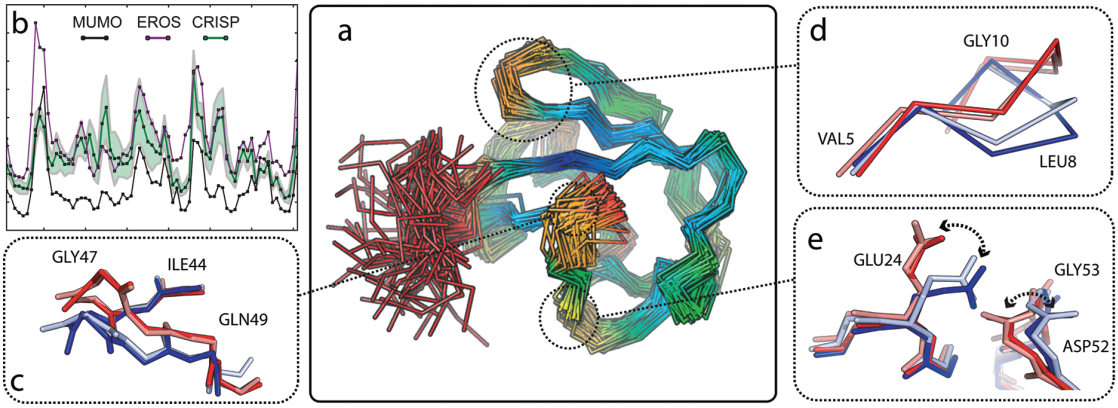


Figure 6. Structural ensemble of ubiquitin obtained with CRISP. (a) Backbone trace of 90 random samples from 10 9 × 10^7 MC-iteration-long simulations using CRISP moves. The residue's color varies from blue (RMSF = 0.5 Å) to green and yellow to red (RMSF > 1.3 Å). (b) RMSF profile from CRISP simulations and from the NMR ensembles EROS and MUMO. (c) The VAL5−LYS11 stretch of the crystal structure 1AAR (dark red) and 2G45 (dark blue). The closest MC samples to the crystal structures are shown in light red and light blue. (d) The ILE44−LEU50 loop of 1AAR (dark red) and 2G45 (dark blue). (e) The ASP52/GLY53 conformational switch. With respect to the crystal structure of 1UBQ (in dark red), this amide-plane flipped state is coupled to the side chain movement of GLU24 and is recurrently found in crystal structures of complexes with deubiquinating enzymes, including 2G45 (in dark blue). Both flipped (light blue) and unflipped (light red) states are explored in our MC sampling.

Table 2. Comparison between the $\overline{\text{MD}}$ Ensemble and MD, SD, CRISP, CRANKSHAFT, and CRA Simulations[a]

	JSD (φ/ψ)	KLD (C_α)
MD	0.07 ± 0.04	174 ± 148
CRISP	0.09 ± 0.05	300 ± 110
SD	0.10 ± 0.06	530 ± 57
CRA	0.13 ± 0.07	1084 ± 115
CRANKSHAFT	0.17 ± 0.09	2265 ± 719

[a] Averages and standard deviations of the JS divergence in φ/ψ space and of the KL divergence of the C_α positions are calculated over 10 runs.

interactions into the derived probability distribution, thus merging backbone and side-chain dynamics into a single move.

One of the goals of our study was to investigate the relative sampling efficiency of Monte Carlo versus molecular dynamics. As has been observed before,[41] our results demonstrate that, in dense environments, molecular dynamics provides a more efficient sampling algorithm than previously described state-of-the-art Monte Carlo methods, presumably due to the high information content provided by the gradient in this scenario. However, this does not seem an inherent limitation of the Monte Carlo method. Our current study indicates, as previously demonstrated for smaller systems,[42,43] that MC can provide the same level of accuracy and efficiency as MD, given that an efficient sampling strategy is used. This result therefore suggests that MC simulations can be reliably employed not only in less dense scenarios (such as intrinsically disordered proteins,[8] protein aggregation,[7] or flexible protein loops[21]) but also for a general *in silico* characterization of flexibility and dynamics of compact molecular systems.

4. MATERIALS AND METHODS

4.1. Molecular Representation. We used a full-atom representation of proteins with fixed bond lengths and flexible dihedral and bond angles. Although flexibility in bond angles is sometimes omitted, it has been shown to increase sampling efficiency.[13,14,44] The peptide dihedral angle ω is included as it is known experimentally to vary at least at the level of bond angles.

4.2. Simulation Setup. The molecular dynamics and stochastic dynamics simulations were conducted using the molecular-modeling package TINKER 5.1.[45] As described in previous studies,[46] stochastic dynamics at constant temperature T = 300 K models the viscous drag of water (frictional coefficient 91 ps^{-1}). Constant temperature T = 300 K molecular dynamics simulations were run using the Beeman integration method. Bond lengths were constrained with the RATTLE algorithm, allowing time steps of 2 fs. MC simulations were conducted using the standard Metropolis-Hastings Monte Carlo scheme at physiological temperature (T = 300 K). Note that several techniques (such as replica exchange[48] or multicanonical ensembles[49]) can be used to enhance the sampling relative to standard MD or Metropolis-Hastings MC simulations. In the present study, for comparative purposes and simplicity, we limited ourselves to direct sampling from the canonical ensemble.

4.3. Force Field. All simulations in this study were conducted using the $OPLS_{aa}$[50] potential in combination with the generalized Born/surface area implicit solvent model GB/SA.[51] Despite the limitations of implicit solvent models, this combination has been widely used and successfully applied to identify the native state of a large set of proteins[52] and for folding simulations.[46,43,47] Note that the CRISP method is not necessarily limited to implicit solvent simulations. Several examples exist of fruitful combinations of Monte Carlo sampling using explicit solvents models.[53,54]

The MC implementation of the $OPLS_{aa}$+GB/SA force field followed that of the Tinker software package and was verified to reproduce the same energy values as this package. The MD and MC results reported in the paper are therefore directly comparable, both in terms of energetics and computational time. We acknowledge that both methods could be optimized further using for instance hardware specific implementations.

4.4. Correlation Times. The MC simulations on Ala_{14} were conducted using the standard Metropolis-Hastings Monte Carlo scheme at physiological temperature (T = 300 K), using only local internal moves (i.e., the N- and C-terminal were kept fixed

during simulation). For CRISP and CRA, the move length was set to five residues. The free parameter λ, which determines the overall size of the CRISP and CRA moves, was tuned for optimality in the context of the correlation times of Ala_{14} (Figure S6). In the CRANKSHAFT move, the number of residues involved in each update was randomly chosen in the range 2−12, as reported in the original description of backrub.[24] Fixed-length CRANKSHAFT moves of length 5 were also attempted but were found to lead to dramatically inferior performance.

4.5. Ubiquitin. All MC simulations on ubiquitin were conducted using the standard Metropolis-Hastings Monte Carlo scheme at T = 300 K. The move-set was composed as follows: 20% local moves, 75% single side-chain moves, and 5% pivot moves. Two different types of single side-chain moves were used: with weight 2/3, samples were drawn from the Dunbrack backbone independent rotamer library[55] (compensating for the bias introduced), while the remaining 1/3 consisted of local side-chain moves (see below). For the pivot moves, new values for the φ and ψ values of a single, randomly chosen residue were drawn from a Gaussian distribution with zero mean and $\sigma = 1°$.

4.6. Side-Chain Sampling. In order to obtain an efficient side-chain sampling, we included a semilocal side-chain move in our move set. Inspired by the biased Gaussian step,[15] this move consists of updating the χ side-chain angles with a constraint toward small displacement of atoms involved as acceptors or donors in hydrogen bonds (Figure S8). This type of move was necessary in order to enable small adjustment of the side-chains without breaking the dense network of noncovalent interactions and was found to greatly facilitate both backbone and side-chain transitions.

It is important to note that all MC methods in our comparison share the same set of Monte Carlo moves for the side chains. It is thus the combination of improved backbone dynamics and efficient side-chain dynamics that gives rise to the increased fluctuations observed with CRISP.

4.7. RMSF Calculations. For each MC/MD simulation, samples were dumped every 2×10^4 MC steps/4 ps and superimposed on the crystal structure 1UBQ, excluding the highly fluctuating terminal residues 71−76. The C_α RMSF of each ensemble was calculated as the root mean squared deviation from the mean position.

4.8. CRANKSHAFT. The CRANKSHAFT move[24] includes an optimized placement of C_β and H_α atoms, which does not fulfill detailed balance and is therefore omitted in our implementation.

4.9. Jensen-Shannon and Kullback−Leibler Divergence. The average Jensen-Shannon divergence $\langle \mathrm{JSD}(\overline{\mathrm{MD}} \| X) \rangle$ between the reference ensemble $\overline{\mathrm{MD}}$ and the ensemble produced by the method X was calculated as

$$\langle \mathrm{JSD}(\overline{\mathrm{MD}} || X) \rangle = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{N} \sum_{j=1}^{N} \mathrm{JSD}(p^j_{\overline{\mathrm{MD}}}(\varphi, \psi) || p^j_{Xi}(\varphi, \psi)) \quad (4)$$

where the index i runs over the 10 simulations, j runs over the residues, and p is the φ,ψ probability distribution estimated using a binning procedure.

The Kullback−Leibler ensemble distance measure was calculated as described in the original reference.[40] In short, assuming the ensembles to be modeled as multivariate normal distributions, it is possible to find a closed-form expression for the KL divergence, which has a direct interpretation in terms of similarity between ensembles. This measure was averaged over 10 runs. All the samples are aligned to the crystal structure 1UBQ prior to the analysis.

4.10. Availability. The CRISP method is implemented as part of the Phaistos software package, freely available under the GNU General Public License v3.0 at sourceforge.net/projects/phaistos.

■ ASSOCIATED CONTENT

Supporting Information. Text S1: Analytical solution for chain closure. Text S2: First order approximation expressing the six postrotational degrees of freedom as a function of the prerotational ones. Text S3: Full expression for the matrix **M** in eq 3. Figure S1: Validation of the first order approximation presented in Text S2. Figure S2: Range of validity of the first order approximation presented in Text S2. Text S4: Outline of the Monte Carlo algorithm. Figure S3: Demonstration of detailed balance for CRISP moves. Figure S4: Angular distribution obtained from Monte Carlo runs using CRISP and from molecular dynamics simulations on alanine 5. Table S1: Comparison between average collective variables calculated using SD and different MC methodologies. Figure S5: Probability distribution and cumulative probability distribution of energy jumps for different local Monte Carlo moves. Figure S6: Dependence of the correlation time over the choice of the parameter λ of eq 3. Figure S7: Jenson-Shannon divergence for the φ/ψ distribution of ubiquitin between MD and SD, CRISP, CRA, and CRANKSHAFT moves. Figure S8: Snapshots from a Monte Carlo simulation on ubiquitin, illustrating a locked side-chain conformation. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sbo@elektro.dtu.dk; wouter.boomsma@thep.lu.se; jfb@elektro.dtu.dk.

Author Contributions

‖W.B. and S.B. contributed equally to this work

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

We thank our colleagues Jes Frellsen, Tim Harder, Kasper Stovgaard, and Mikael Borg for useful discussions and Jacobus J. Boomsma and Kresten Lindorff-Larsen for valuable comments and suggestions. S.B is supported by Radiometer. W.B and K.E.J are funded by the Danish Council for Independent Research (FNU, 272-08-0315 and FTP, 274-08-0124, respectively). C.A. and T.H. acknowledge the Danish Program Commission on Nanoscience (NaBiIT, 2106-06-0009).

■ REFERENCES

(1) Eisenmesser, E.; Millet, O.; Labeikovsky, W.; Korzhnev, D.; Wolf-Watz, M.; Bosco, D.; Skalicky, J.; Kay, L.; Kern, D. *Nature* **2005**, *438*, 117–121.

(2) Chiti, F.; Dobson, C. *Nat. Chem. Biol.* **2008**, *5*, 15–22.

(3) Nevo, R.; Stroh, C.; Kienberger, F.; Kaftan, D.; Brumfeld, V.; Elbaum, M.; Reich, Z.; Hinterdorfer, P. *Nat. Struct. Mol. Biol.* **2003**, *10*, 553–557.

(4) Boehr, D.; Nussinov, R.; Wright, P. *Nat. Chem. Biol.* **2009**, *5*, 789–796.
(5) Ponder, J.; Case, D. *Adv. Protein Chem.* **2003**, *66*, 27–85.
(6) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.
(7) Li, D.; Mohanty, S.; Irbäck, A.; Huo, S. *PLoS Comput. Biol.* **2008**, *4*, e1000238.
(8) Mao, A.; Crick, S.; Vitalis, A.; Chicoine, C.; Pappu, R. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183.
(9) Gō, N.; Scheraga, H. *Macromolecules* **1970**, *3*, 178–187.
(10) Dodd, L.; Boone, T.; Theodorou, D. *Mol. Phys.* **1993**, *78*, 961–996.
(11) Hoffmann, D.; Knapp, E. *Eur. Biophys. J.* **1996**, *24*, 387–403.
(12) Dinner, A. *J. Comput. Chem.* **2000**, *21*, 1132–1144.
(13) Ulmschneider, J.; Jorgensen, W. *J. Chem. Phys.* **2003**, *118*, 4261–4271.
(14) Bruccoleri, R.; Karplus, M. *Macromolecules* **1985**, *18*, 2767–2773.
(15) Favrin, G.; Irbäck, A.; Sjunnesson, F. *J. Chem. Phys.* **2001**, *114*, 8154–8158.
(16) Frenkel, D.; Mooij, G.; Smit, B. *J. Phys.: Condens. Matter* **1992**, *4*, 3053–3076.
(17) Escobedo, F. J.; de Pablo, J. J. *J. Chem. Phys.* **1995**, *102*, 2636–2652.
(18) Vendruscolo, M. *J. Chem. Phys.* **1997**, *106*, 2970–2976.
(19) Chen, Z.; Escobedo, F. *J. Chem. Phys.* **2000**, *113*, 11382–11392.
(20) Coutsias, E.; Seok, C.; Jacobson, M.; Dill, K. *J. Comput. Chem.* **2004**, *25*, 510–528.
(21) Nilmeier, J.; Hua, L.; Coutsias, E.; Jacobson, M. *J. Chem. Theory Comput.* **2011**, *7*, 1564–1574.
(22) Betancourt, M. *J. Chem. Phys.* **2005**, *123*, 174905–174905−07.
(23) Davis, I.; Arendall, W., III; Richardson, D.; Richardson, J. *Structure* **2006**, *14*, 265–274.
(24) Smith, C.; Kortemme, T. *J. Mol. Biol.* **2008**, *380*, 742–756.
(25) Lauck, F.; Smith, C.; Friedland, G.; Humphris, E.; Kortemme, T. *Nucleic Acids Res.* **2010**, *38*, W569–W575.
(26) Engh, R.; Huber, R. *Acta Crystallogr., Sect. A* **1991**, *47*, 392–400.
(27) Frenkel, D.; Smit, B. Appendix D. In *Understanding molecular simulation: from algorithms to applications*, 2nd ed.; Academic Press: San Diego, CA, 2002; pp 525−532.
(28) Hershko, A.; Ciechanover, A. *Annu. Rev. Biochem.* **1998**, *67*, 425–479.
(29) Hicke, L.; Schubert, H.; Hill, C. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 610–621.
(30) Tjandra, N.; Feller, S.; Pastor, R.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562–12566.
(31) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.
(32) Chou, J.; David, A.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 8959–8966.
(33) Lange, O.; Lakomek, N.-A.; Fares, C.; Schroder, G.; Walter, K.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. *Science* **2008**, *320*, 1471–1475.
(34) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M.; Dror, R.; Klepeis, J.; Arkin, I.; Jensen, M.; Xu, H.; Trbovic, N.; Friesner, R. *J. Phys. Chem. B* **2008**, *112*, 6155–6158.
(35) Richter, B.; Gsponer, J; Varnai, P.; Salvatella, X; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.
(36) Lindorff-Larsen, K.; Best, R.; DePristo, M.; Dobson, C.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(37) Nederveen, A.; Bonvin, A. *J. Chem. Theory Comput.* **2005**, *1*, 363–374.
(38) Paterlini, M.; Ferguson, D. *Chem. Phys.* **1998**, *236*, 243–252.
(39) Huang, K. Y.; Amodeo, G. A.; Tong, L.; McDermott, A. *Protein Sci.* **2011**, *20*, 630–639.
(40) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS One* **2009**, *4*, e4203.
(41) Yamashita, H.; Endo, S.; Wako, H.; Kidera, A. *Chem. Phys. Lett.* **2001**, *342*, 382–386.
(42) Jorgensen, W.; Tirado-Rives, J. *J. Phys. Chem.* **1996**, *100*, 14508–14513.
(43) Ulmschneider, J.; Ulmschneider, M.; Di Nola, A. *J. Phys. Chem. B* **2006**, *110*, 16733–16742.
(44) Karplus, M. *Methods Enzymol.* **1986**, *131*, 283–307.
(45) Ponder, J.; Richards, F. *J. Am. Chem. Soc.* **1987**, *8*, 1016–1024.
(46) Snow, C.; Qiu, L.; Du, D.; Gai, F.; Hagen, S.; Pande, V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077–4082.
(47) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.
(48) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(49) Ferkinghoff-Borg, J. *Eur. Phys. J. B* **2002**, *29*, 481–484.
(50) Jorgensen, D. S.; Maxwell, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *18*, 11225–11236.
(51) Di Qiu Shenkin, F. P.; Hollinger, P. S.; Still, W. C. *J. Phys. Chem. A.* **1997**, *101*, 3005–3014.
(52) Chopraa, C. M.; Summab, G.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20239–20244.
(53) Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D. *Proteins* **2005**, *58*, 893–904.
(54) Schymkowitz, J.; Rousseau, F.; Martins, I.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10147.
(55) Dunbrack, R. L.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.

Subtle Monte Carlo updates in dense molecular systems

Sandro Bottaro, Wouter Boomsma, Kristoffer E. Johansson, Christian Andreetta, Thomas Hamelryck, and Jesper Ferkinghoff-Borg

Text S1 - Analytical solution for chain closure

Figure 1 illustrates the degrees of freedom involved in the post-rotation. The leftmost C, N and C_α atoms are the last positions that are affected by the pre-rotation, and will remain fixed during post-rotation. By construction, the positions of the rightmost N and C_α and C atoms as well as all the bond lengths and the dihedral angle ω_p should be unaffected by the local move. Only the position $\vec{r}_2$ of the central C atom will be updated during post-rotation, resulting in new values for the dihedral angles χ_1, χ_3, χ_6, and bond angles χ_2, χ_4, χ_5 (in blue, Fig. 1).

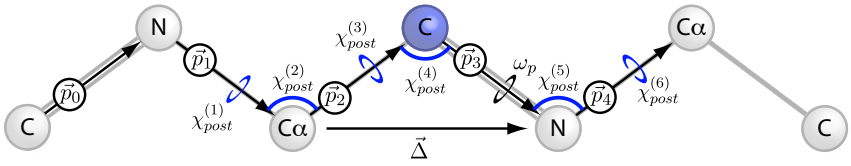


Figure 1. The post-rotation step. This step proceeds by calculating the position of the central C atom (in blue), from which all bond vectors $\vec{p}_0 \dots \vec{p}_4$ can be determined, resulting in new values for the 6 post-rotational degrees of freedom $\chi^{(1)}_{post} \dots \chi^{(6)}_{post}$. The peptide dihedral angle ω_p does not vary during post-rotation.

Let i represent the atom number along the backbone, and $\vec{r}_i$ be the corresponding position vectors relative to the origin of the global coordinate system. By assumption, the lengths p_i of all bond vectors $\vec{p}_i = \vec{r}_i - \vec{r}_{i-1}$ are known. Since $\vec{r}_1$ and $\vec{r}_3$ are known, the vector $\vec{\Delta} = \vec{r}_3 - \vec{r}_1$, is also known (see Fig. 1). In the following we determine $\vec{p}_3$ from $\vec{\Delta}$ and ω_p and use $\vec{r}_2 = \vec{r}_3 - \vec{p}_3$ to obtain the desired position of the C-atom. In order to determine $\vec{p}_3$ it is convenient to define the orthonormal basis $\{\hat{e}_i\}$

$$\hat{e}_1 = \frac{\vec{p}_2 + \vec{p}_3}{|\vec{p}_2 + \vec{p}_3|} = \frac{\vec{\Delta}}{\Delta}, \quad \hat{e}_2 = \frac{\vec{p}_4 - (\vec{p}_4 \cdot \hat{e}_1)\hat{e}_1}{|\vec{p}_4 - (\vec{p}_4 \cdot \hat{e}_1)|}, \quad \hat{e}_3 = \hat{e}_1 \times \hat{e}_2. \tag{1}$$

Applying the law of cosines, the projection of $\vec{p}_3$ on $\hat{e}_1$ is given by

$$\vec{p}_3 \cdot \hat{e}_1 = p_{3,1} = \frac{\Delta^2 + p_3^2 - p_2^2}{2\Delta}. \tag{2}$$

Since ω_p is fixed we can use the definition of the dihedral angle

$$\cos(\omega_P) = \frac{(\vec{p}_2 \times \vec{p}_3) \cdot (\vec{p}_3 \times \vec{p}_4)}{|(\vec{p}_2 \times \vec{p}_3)||(\vec{p}_3 \times \vec{p}_4)|} = \frac{(\hat{e}_1 \times \vec{p}_3) \cdot (\vec{p}_3 \times \vec{p}_4)}{|(\hat{e}_1 \times \vec{p}_3)||(\vec{p}_3 \times \vec{p}_4)|} \tag{3}$$

to obtain a quadratic expression for $\vec{p}_3 \cdot \vec{p}_4$. Squaring Eq. (3) and applying the vectorial identity $(\vec{a} \times \vec{b}) \cdot (\vec{c} \times \vec{d}) = (\vec{a} \cdot \vec{c})(\vec{b} \cdot \vec{d}) - (\vec{a} \cdot \vec{d})(\vec{c} \cdot \vec{b})$ we obtain

$$(p_{3,1}^2 + \tilde{p}_{3,\perp}^2)[\vec{p}_3 \cdot \vec{p}_4]^2 - 2p_3^2 p_{3,1} p_{4,1}[\vec{p}_3 \cdot \vec{p}_4] + p_3^2(p_3^2 p_{4,1}^2 - p_4^2 \tilde{p}_{3,\perp}^2) = 0 \tag{4}$$

where $\vec{p}_{3,\perp}^{\,2} = (p_3^2 - p_{3,1}^2)\cos^2(\omega_P)$ and $p_{4,1} = \vec{p}_4 \cdot \hat{e}_1$. Eq. (4) provides two solutions. To avoid large conformational changes we choose the solution for which the scalar product

$$\vec{p}_3 \cdot \hat{e}_2 = p_{3,2} = \frac{\vec{p}_3 \cdot \vec{p}_4 - p_{3,1}p_{4,1}}{|\vec{p}_4 \cdot \hat{e}_2|} \tag{5}$$

deviates least from the same quantity calculated from the original structure. Finally, the component of $\vec{p}_3$ along $\hat{e}_3$ is given by

$$\vec{p}_3 \cdot \hat{e}_3 = p_{3,3} = \pm\sqrt{p_3^2 - p_{3,1}^2 - p_{3,2}^2}. \tag{6}$$

Once again, the sign of the solution is chosen according to the original structure. Eq. (2), (5) and (6) give the three components of $\vec{p}_3$ with respect to the known ortonormal basis, Eq. (1) and which allows us to position the C-atom, $\vec{r}_2$. From here, all post-rotational degrees of freedom, $\{\chi_i\}$, can then be obtained.

We emphasize that in the calculations carried out so far we have fixed the peptide dihedral angle ω_p, in order to make our kinetic approach equally applicable to cases where all ω-angles are kept fixed. From a mathematical point of view, however, it is easier to fix the bond angle χ_5, releasing the degree of freedom ω_p. In this case the scalar product $\vec{p}_3 \cdot \vec{p}_4$ is given by the original structure and the projection $p_{3,2}$ in Eq. 5 is readily determined.

Text S2 - First order approximation

The derived analytical solution allows us to express the six post-rotational degrees of freedom $\bar{\chi}_{post} = (\chi_{post}^{(1)} \dots \chi_{post}^{(6)})$ as a function of the pre-rotational ones $\bar{\chi}_{pre} = (\chi_{pre}^{(1)} \dots \chi_{pre}^{(n-6)})$. The change of each post rotational degrees of freedom, $\delta\chi_{post}^{(k)}$ upon the pre-rotation $\delta\bar{\chi}_{pre}$ can be evaluated to first order as

$$\delta\chi_{post}^{(k)} = \sum_{j=1}^{n-6} \frac{\partial\chi_{post}^{(k)}}{\partial\chi_{pre}^{(j)}}\delta\chi_{pre}^{(j)} = \sum_{j=1}^{n-6} S_{k,j}\delta\chi_{pre}^{(j)} \tag{7}$$

where the index j runs over the $n-6$ pre-rotational degrees of freedom and S is a $6 \times (n-6)$ matrix. This matrix is most easily evaluated using

$$S_{k,j} = \frac{\partial\chi_{post}^{(k)}}{\partial\chi_{pre}^{(j)}} = \sum_{i=0}^{3} \frac{\partial\chi_{post}^{(k)}}{\partial\vec{p}_i} \cdot \frac{\partial\vec{p}_i}{\partial\chi_{pre}^{(j)}} = \sum_{i=0}^{3} \vec{Z}_{k,i} \cdot \vec{\Lambda}_{i,j} \tag{8}$$

Here, the index i runs over the four bond vectors $\vec{p}_0 \dots \vec{p}_3$. For convenience and ease of reference, we have introduced the matrices $\vec{\mathbf{Z}}$ and $\vec{\mathbf{\Lambda}}$ with vectorial elements, $\vec{Z}_{k,i} = \frac{\partial\chi_{post}^{(k)}}{\partial\vec{p}_i}$ and $\vec{\Lambda}_{i,j} = \frac{\partial\vec{p}_i}{\partial\chi_{pre}^{(j)}}$. The multiplication between these matrix elements is understood as a scalar product. We shall derive each of these matrices in turn.

Derivation of $\vec{\mathbf{Z}}$

To calculate $\vec{Z}_{k,i} = \partial\chi_{post}^{(k)}/\partial\vec{p}_i$, we destinguish between the case where k refers to a bond (b) angle and where it refers to a dihedral (d) angle. Furthermore, we shall refer to $\vec{p}_i$ using a relative indexation, $\vec{p}_-$, $\vec{p}$ and $\vec{p}_+$, according to how $\vec{p}_i$ relates to $\chi^{(k)}$ (Fig. 2). Consequently, if χ_k is a bond angle then $\vec{Z}_{k,i} = \vec{0}$ when $\vec{p}_i$ does not equal either $\vec{p}_-$ or $\vec{p}$. Similarly, $\vec{Z}_{k,i} = \vec{0}$ when χ_k is a dihedral angle and $\vec{p}_i$ does not equal either $\vec{p}_-, \vec{p}$ or $\vec{p}_+$. When χ_k is bond angle, χ_b, we derive from

$$\vec{p}_- \cdot \vec{p} = -\cos(\chi_b)p_-p \tag{9}$$

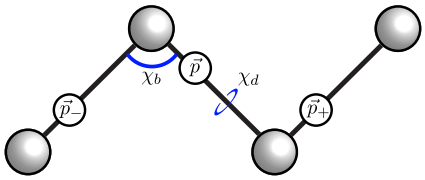


Figure 2. Definition of the notational relation between angles, χ and bond vectors $\vec{p}$. The bond angle, χ_b is defined from the two consecutive bond vectors $\vec{p}_-$ and $\vec{p}$. The dihedral angle, χ_d, is defined from the three consecutive bond vectors $\vec{p}_-$, $\vec{p}$ and $\vec{p}_+$.

and

$$\delta\cos(\chi) = -\sin(\chi)\delta\chi \tag{10}$$

the following expressions

$$\frac{\partial\chi_b}{\partial\vec{p}_-} = -\frac{1}{\sin(\chi_b)}\frac{\vec{p}}{p_- p} \quad ; \quad \frac{\partial\chi_b}{\partial\vec{p}} = -\frac{1}{\sin(\chi_b)}\frac{\vec{p}_-}{p_- p}. \tag{11}$$

We note that these two expressions are always well-defined since $\sin(\chi_b)$ is never close to zero for a bond angle. However, this is not necessarily true for dihedral angles, χ_d. To ensure that the expression for $\vec{Z}_{k,i}$ is always evaluable in this case as well, irrespective of the value of χ_d, we base the derivation on two alternative formulas. For $\chi_d \in [\frac{\pi}{4}, \frac{3}{4}\pi[\cup [-\frac{3}{4}\pi, -\frac{1}{4}\pi[$, we use the definition

$$\cos(\chi_d) = \vec{b}_- \cdot \vec{b}_+ \tag{12}$$

whereas for $\chi_d \in [\frac{3}{4}\pi, \frac{5}{4}\pi[\cup [-\frac{1}{4}\pi, \frac{1}{4}\pi[$, we use

$$\sin(\chi_d) = \frac{\vec{p}}{p} \cdot (\vec{b}_- \times \vec{b}_+). \tag{13}$$

Here, $\vec{b}_- = \frac{\vec{p}_- \times \vec{p}}{|\vec{p}_- \times \vec{p}|}$ and $\vec{b}_+ = \frac{\vec{p} \times \vec{p}_+}{|\vec{p} \times \vec{p}_+|}$.

Differentiating the cosine-version of χ_d, Eq. (12), with respect to $\vec{p}_i$ leads to

$$\frac{\partial\cos(\chi_d)}{\partial\vec{p}_i} = \frac{\partial\vec{b}_+}{\partial\vec{p}_i} \cdot \vec{b}_- + \frac{\partial\vec{b}_-}{\partial\vec{p}_i} \cdot \vec{b}_+. \tag{14}$$

To evaluate the two terms, we observe that the differential of any vector $\vec{b} = \frac{\vec{v}}{v}$ is given by

$$\delta\vec{b} = \frac{\delta\vec{v} - (\vec{b} \cdot \delta\vec{v})\vec{b}}{v}. \tag{15}$$

Furthermore, for all $\vec{p}$, $\vec{q}$, $\vec{r}$ [1]

$$\frac{\partial(\vec{p} \times \vec{q})}{\partial\vec{p}} \cdot \vec{r} = \vec{q} \times \vec{r}, \qquad \frac{\partial(\vec{p} \times \vec{q})}{\partial\vec{q}} \cdot \vec{r} = \vec{r} \times \vec{p}. \tag{16}$$

[1] This equation is demonstrated by rewriting in tensor notation

$$[\frac{\partial(\vec{p} \times \vec{q})}{\partial\vec{p}} \cdot \vec{r}]_\mu = \frac{\partial\epsilon_{\alpha\beta\gamma}p_\beta q_\gamma}{\partial p_\mu} r_\alpha = \epsilon_{\alpha\mu\gamma} q_\mu r_\alpha = [\vec{q} \times \vec{r}]_\mu,$$

where ϵ is the Levi-Civita symbol.

From Eq. (15) and Eq. (16) we can easily evaluate Eq. (14). For instance, for any vector $\vec{q}$

$$\frac{\partial \vec{b}_-}{\partial \vec{p}_-} \cdot \vec{q} = \frac{\partial}{\partial \vec{p}_-} \left(\frac{\vec{p}_- \times \vec{p}}{|\vec{p}_- \times \vec{p}|} \right) \cdot \vec{q} = \frac{(\vec{p} \times \vec{q}) - (\vec{p} \times \vec{b_-})(\vec{b}_- \cdot \vec{q})}{|\vec{p}_- \times \vec{p}|}. \tag{17}$$

Consequently, for the three cases of interest $(-, \cdot, +)$, Eq. (14) becomes

$$\begin{aligned}
\frac{\partial \cos(\chi_d)}{\partial \vec{p}_-} &= k_1 \left[(\vec{p} \times \vec{b}_+) - \cos(\chi_d)(\vec{p} \times \vec{b}_-) \right] \\
\frac{\partial \cos(\chi_d)}{\partial \vec{p}} &= k_1 \left[(\vec{b}_+ \times \vec{p}_-) + \cos(\chi_d)(\vec{p}_- \times \vec{b}_-) \right] + k_2 \left[(\vec{p}_+ \times \vec{b}_-) - \cos(\chi_d)(\vec{p}_+ \times \vec{b}_+) \right] \\
\frac{\partial \cos(\chi_d)}{\partial \vec{p}_+} &= k_2 \left[(\vec{b}_- \times \vec{p}) + \cos(\chi_d)(\vec{p} \times \vec{b}_+) \right]
\end{aligned} \tag{18}$$

where we have defined

$$k_1 = \frac{1}{|\vec{p}_- \times \vec{p}|}, \qquad k_2 = \frac{1}{|\vec{p} \times \vec{p}_+|}.$$

Finally $\vec{Z}_{k,i}$ is obtained using

$$\frac{\partial \chi_d}{\partial \vec{p}_i} = -\frac{1}{\sin(\chi_d)} \frac{\partial \cos(\chi_d)}{\partial \vec{p}_i}.$$

When $\chi_d \in [\frac{3}{4}\pi, \frac{5}{4}\pi[\cup [-\frac{1}{4}\pi, \frac{1}{4}\pi[$, we determine the matrix elements from the differential of Eq. (13)

$$\begin{aligned}
\frac{\partial \sin(\chi_d)}{\partial \vec{p}_i} &= \frac{\partial}{\partial \vec{p}_i} \left(\frac{\vec{p}}{p} \right) \cdot (\vec{b}_- \times \vec{b}_+) + \frac{\vec{p}}{p} \cdot \left(\frac{\partial \vec{b}_-}{\partial \vec{p}_i} \times \vec{b}_+ + \vec{b}_- \times \frac{\partial \vec{b}_+}{\partial \vec{p}_i} \right) \\
&= \frac{\partial}{\partial \vec{p}_i} \left(\frac{\vec{p}}{p} \right) \cdot (\vec{b}_- \times \vec{b}_+) + \frac{\partial \vec{b}_-}{\partial \vec{p}_i} \cdot \left(\vec{b}_+ \times \frac{\vec{p}}{p} \right) - \frac{\partial \vec{b}_+}{\partial \vec{p}_i} \cdot \left(\vec{b}_- \times \frac{\vec{p}}{p} \right)
\end{aligned} \tag{19}$$

Therefore, by applying Eq. (17) and the vector identities

$$\vec{a} \cdot (\vec{b} \times \vec{c}) = (\vec{a} \times \vec{b}) \cdot \vec{c} = (\vec{c} \times \vec{a}) \cdot \vec{b}, \qquad (\vec{a} \times \vec{b}) \times \vec{c} = (\vec{a} \cdot \vec{c})\vec{b} - (\vec{b} \cdot \vec{c})\vec{a}, \qquad \vec{a} \times (\vec{b} \times \vec{c}) = (\vec{a} \cdot \vec{c})\vec{b} - (\vec{a} \cdot \vec{b})\vec{c}$$

we obtain

$$\begin{aligned}
\frac{\partial \sin(\chi_d)}{\partial \vec{p}_-} &= k_1 \left[p\vec{b}_+ - \sin(\chi_d)(\vec{p} \times \vec{b}_-) \right] \\
\frac{\partial \sin(\chi_d)}{\partial \vec{p}} &= \frac{\vec{b}_- \times \vec{b}_+}{p} + k_1 \left[\frac{(\vec{b}_+ \times \vec{p}) \times \vec{p}_-}{p} + \sin(\chi_d)(\vec{p}_- \times \vec{b}_-) \right] + k_2 \left[\frac{(\vec{b}_- \times \vec{p}) \times \vec{p}_+}{p} - \sin(\chi_d)(\vec{p}_+ \times \vec{b}_+) \right] \\
\frac{\partial \sin(\chi_d)}{\partial \vec{p}_+} &= k_2 \left[p\vec{b}_- + \sin(\chi_d)(\vec{p} \times \vec{b}_+) \right]
\end{aligned} \tag{20}$$

The matrix elements, $\vec{Z}_{k,i}$, for $\chi_d \in [\frac{3}{4}\pi, \frac{5}{4}\pi[\cup [-\frac{1}{4}\pi, \frac{1}{4}\pi[$, are then finally obtained as

$$\frac{\partial \chi_d}{\partial \vec{p}_i} = \frac{1}{\cos(\chi_d)} \frac{\partial \sin(\chi_d)}{\partial \vec{p}_i}. \tag{21}$$

Derivation of $\vec{\Lambda}$

The matrix elements, $\vec{\Lambda}_{i,j} = \frac{\partial \vec{p}_i}{\partial \chi_{pre}^{(j)}}$ can be found directly when $i = 0, 1$ from the properties of rotation

matrices. Let $\hat{n}_j$ be the normalized vector associated with the j'th prerotational degree of freedom $\chi^{(j)}_{pre}$, defined as

$$\hat{n}_j = \begin{cases} \frac{\vec{p}^{(j)}_- \times \vec{p}^{(j)}}{|\vec{p}^{(j)}_- \times \vec{p}^{(j)}|} & \text{if } \chi^{(j)}_{pre} \text{ is bond angle} \\ \frac{\vec{p}^{(j)}}{p^{(j)}} & \text{if } \chi^{(j)}_{pre} \text{ is dihedral angle} \end{cases}$$

where $\vec{p}^{(j)}_-$ and $\vec{p}^{(j)}$ are bond vectors associated with the j'th prerotational degree of freedom, c.f. Fig. 2. Then

$$\frac{\partial \vec{p}_i}{\partial \chi^{(j)}_{pre}} = \hat{n}_j \times \vec{p}_i, \quad \text{for } i = 0, 1. \tag{22}$$

To determine the matrix elements for $\vec{\Lambda}_{i,j}$ when $i = 2, 3$, we shall take the following approach. First, we determine the tensorial relation between $\delta\vec{p}_3$ and $\delta\vec{r}_1$ from the analytical chain-closure solution. Let this relation be written as $\delta\vec{p}_3 = \Pi_3 \cdot \delta\vec{r}_1$, where Π_3 is a rank two tensor[2]. Secondly, we apply the known relation between $\delta\vec{r}_1$ and $\delta\chi_j$, which is given by an expression similar to Eq. (22). Combining these results will yield the matrix elements $\vec{\Lambda}_{i=3,j}$. The remaining matrix-elements, $\vec{\Lambda}_{i=2,j}$, are finally obtained from $\vec{\Lambda}_{i=3,j}$, using the relation $\delta(\vec{p}_2 + \vec{p}_3) = \delta\vec{\Delta} = -\delta\vec{r}_1$ which implies

$$\delta\vec{p}_2 = -\delta\vec{p}_3 - \delta\vec{r}_1 = -(\Pi_3 + \mathbb{I}) \cdot \delta\vec{r}_1, \tag{23}$$

where $\mathbb{I}$ is the identity tensor.

To determine Π_3 we begin by projecting the variation of the vector $\vec{p}_3$ on the orthonormal basis defined in Eq. (1)

$$\delta\vec{p}_3 = \sum_{i=1}^{3} \delta((\vec{p}_3 \cdot \hat{e}_i)\hat{e}_i) = \sum_{i=1}^{3} \delta(p_{3,i}\hat{e}_i) = \sum_{i=1}^{3} p_{3,i}\delta\hat{e}_i + \sum_{i=1}^{3} \hat{e}_i \delta p_{3,i} \tag{24}$$

We calculate each of these terms in turn. The differential of $\hat{e}_1$ is given by

$$\delta\hat{e}_1 = \delta\left(\frac{\vec{\Delta}}{\Delta}\right) = \frac{1}{\Delta}(\delta\vec{\Delta} - (\hat{e}_1 \cdot \delta\vec{\Delta})\hat{e}_1) = (\mathbb{I} - \hat{e}_1\hat{e}_1) \cdot \frac{\delta\vec{\Delta}}{\Delta}.$$

Here, $\hat{e}_1\hat{e}_1$ represents the outer product between $\hat{e}_1$ and $\hat{e}_1$[3]. Defining also $\delta\vec{\Delta}' = \frac{\delta\vec{\Delta}}{\Delta}$, we write

$$\delta\hat{e}_1 = (\hat{e}_2\hat{e}_2 + \hat{e}_3\hat{e}_3) \cdot \delta\vec{\Delta}', \tag{25}$$

where $\mathbb{I} = \sum_{n=1}^{3} \hat{e}_n\hat{e}_n$ has been used. Using Eq. (15) the differential of the second basis vector becomes

$$\begin{aligned} \delta\hat{e}_2 &= \delta\left(\frac{\vec{p}_4 - p_{4,1}\hat{e}_1}{|\vec{p}_4 - p_{4,1}\hat{e}_1|}\right) = \frac{1}{p_{4,2}}\left[\delta(\vec{p}_4 - p_{4,1}\hat{e}_1) - (\delta(\vec{p}_4 - p_{4,1}\hat{e}_1) \cdot \hat{e}_2)\hat{e}_2\right] \\ &= \frac{1}{p_{4,2}}\left[-\delta(p_{4,1}\hat{e}_1) + (\delta(p_{4,1}\hat{e}_1) \cdot \hat{e}_2)\hat{e}_2\right], \end{aligned} \tag{26}$$

where we have used the fact that $\vec{p}_4$ remains constant, $\delta\vec{p}_4 = 0$. Since $p_{4,3} = 0$ the variation of $p_{4,1}$ becomes

$$\delta p_{4,1} = \vec{p}_4 \cdot \delta\hat{e}_1 = p_{4,2}(\hat{e}_2 \cdot \delta\vec{\Delta}')., \tag{27}$$

[2] A rank two tensor corresponds uniquely to a 3×3 matrix once a basis system is defined.

[3] In the rest of this text the multiplication between two vectors, $\vec{a}\vec{b}$, is understood as an outer product, unless a different product ($\cdot$ or $\times$) is explicitly specified. We remind the reader that the outer product is a rank two tensor defined from the relations $(\vec{a}\vec{b}) \cdot \vec{c} = \vec{a}(\vec{b} \cdot \vec{c})$ and $\vec{c} \cdot (\vec{a}\vec{b}) = (\vec{c} \cdot \vec{a})\vec{b}$. Consequently, if a_i and b_j are the components of $\vec{a}$ and $\vec{b}$ with respect to a given basis then $\vec{a}\vec{b}$ can be represented as a matrix, M, with elements $M_{ij} = a_i b_j$.

where Eq. (25) has been used. Using the linearity of differentials and inserting Eq. (25) and Eq. (27) in Eq. (26) we obtain

$$\delta\hat{e}_2 = \left(\frac{p_{4,1}}{p_{4,2}}\hat{e}_2\hat{e}_2 - \hat{e}_1\hat{e}_2 - \frac{p_{4,1}}{p_{4,2}}\hat{e}_2\hat{e}_2 - \frac{p_{4,1}}{p_{4,2}}\hat{e}_3\hat{e}_3\right)\cdot\delta\vec{\Delta}' = -(\hat{e}_1\hat{e}_2 + \frac{p_{4,1}}{p_{4,2}}\hat{e}_3\hat{e}_3)\cdot\delta\vec{\Delta}' \tag{28}$$

Consequently,

$$\begin{aligned}\delta\hat{e}_3 &= \delta(\hat{e}_1\times\hat{e}_2) = \hat{e}_1\times\delta\hat{e}_2 - \hat{e}_2\times\delta\hat{e}_1 = -\hat{e}_1\times\left(\hat{e}_1\hat{e}_2 + \frac{p_{4,1}}{p_{4,2}}\hat{e}_3\hat{e}_3\right)\cdot\delta\vec{\Delta} - \hat{e}_2\times(\hat{e}_2\hat{e}_2 + \hat{e}_3\hat{e}_3)\cdot\delta\vec{\Delta}' \\ &= \left(\frac{p_{4,1}}{p_{4,2}}\hat{e}_2\hat{e}_3 - \hat{e}_1\hat{e}_3\right)\cdot\delta\vec{\Delta}'. \end{aligned}\tag{29}$$

This concludes the first summation on the right hand side of Eq. (24). The first term in the second summation is directly evaluable from the law of cosines

$$\begin{aligned}\delta p_{3,1} &= \delta\left(\frac{\Delta^2 + p_3^2 - p_2^2}{2\Delta}\right) = (1 - \frac{p_{3,1}}{\Delta})\delta\Delta = (1 - \frac{p_{3,1}}{\Delta})\frac{\vec{\Delta}\cdot\delta\vec{\Delta}}{\Delta} \\ &= (\Delta - p_{3,1})(\hat{e}_1\cdot\delta\vec{\Delta}') = p_{2,1}(\hat{e}_1\cdot\delta\vec{\Delta}'). \end{aligned}\tag{30}$$

The second term, $\delta p_{3,2}$, is found from

$$\delta p_{3,2} = \delta\left(\vec{p}_3\cdot\frac{\vec{p}_4 - p_{4,1}\hat{e}_1}{p_{4,2}}\right) \tag{31}$$

$$= \frac{1}{p_{4,2}}\left[\delta(\vec{p}_3\cdot\vec{p}_4) - \delta(p_{3,1}p_{4,1}) - p_{3,2}\delta p_{4,2}\right]. \tag{32}$$

This expression involves $\delta p_{3,1}$, $\delta p_{4,1}$, $\delta p_{4,2}$ and $\delta(\vec{p}_3\cdot\vec{p}_4)$. The first two differentials are given by Eq. (30) and Eq. (27), respectively. The third differential is obtained from Eq. (28)

$$\delta p_{4,2} = \vec{p}_4\cdot\delta\hat{e}_2 = -p_{4,1}(\hat{e}_2\cdot\delta\vec{\Delta}'). \tag{33}$$

Finally, $\delta(\vec{p}_3\cdot\vec{p}_4)$ is related to $\delta p_{3,1}$ and $\delta p_{4,1}$ through the quadratic formula in Eq. (4). Taking the full differential of Eq. (4) leads to

$$\delta(\vec{p}_3\cdot\vec{p}_4) = -\frac{\delta a(\vec{p}_3\cdot\vec{p}_4)^2 + \delta b(\vec{p}_3\cdot\vec{p}_4) + \delta c}{2a(\vec{p}_3\cdot\vec{p}_4) + b} \tag{34}$$

where a, b, c are the coefficients

$$\begin{aligned} a &= p_3^2\cos^2(\omega_p) + p_{3,1}^2\sin^2(\omega_p) \\ b &= -2p_3^2p_{3,1}p_{4,1} \\ c &= p_3^2(p_3^2p_{4,1}^2 - (p_3^2 - p_{3,1}^2)p_4^2\cos^2(\omega_p)) \end{aligned}$$

and

$$\begin{aligned} \delta a &= 2p_{3,1}\sin^2(\omega_p)\delta p_{3,1} \\ \delta b &= -2p_3^2(p_{4,1}\delta p_{3,1} + p_{3,1}\delta p_{4,1}) \\ \delta c &= 2p_3^2(p_3^2p_{4,1}\delta p_{4,1} + p_4^2p_{3,1}\cos^2(\omega_p)\delta p_{3,1}) \end{aligned}$$

Consequently,

$$\begin{aligned}\delta(\vec{p}_3\cdot\vec{p}_4) = &- \frac{(\vec{p}_3\cdot\vec{p}_4)^2 p_{3,1}\sin^2(\omega_p) - (\vec{p}_3\cdot\vec{p}_4)p_3^2 p_{4,1}}{D}\delta p_{3,1} - \frac{p_3^2 p_4^2 p_{3,1}\cos^2(\omega_p)}{D}\delta p_{3,1} + \\ &- \frac{p_3^4 p_{4,1} - (\vec{p}_3\cdot\vec{p}_4)p_{3,1}p_3^2}{D}\delta p_{4,1} = A\delta p_{3,1} + B\delta p_{4,1}, \end{aligned} \tag{35}$$

where we have defined $D = p_3^2(\vec{p}_3\cdot\vec{p}_4)\cos^2(\omega_p) - p_{3,1}p_{4,1}p_3^2 + p_{3,1}^2\vec{p}_3\cdot\vec{p}_4\sin^2(\omega_p)$. Note that if one chooses to fix the bond angle (χ_5 in Fig. 1) instead of the dihedral angle (ω_p in Fig. 1), $\delta(\vec{p}_3\cdot\vec{p}_4)$ vanishes. Inserting equations (30),(27), (33),(35) in Eq. (32) we get

$$\begin{aligned}\delta p_{3,2} &= \frac{p_{2,1}}{p_{4,2}}(A - p_{4,1})(\hat{e}_1\cdot\delta\vec{\Delta}') + \left(\frac{p_{3,2}p_{4,1} - p_{3,1}p_{4,2}}{p_{4,2}} + B\right)(\hat{e}_2\cdot\delta\vec{\Delta}') \\ &= \tilde{A}(\hat{e}_1\cdot\delta\vec{\Delta}') + \tilde{B}(\hat{e}_2\cdot\delta\vec{\Delta}'). \end{aligned} \tag{36}$$

Returning to Eq. (24), the only remaining differential is $\delta p_{3,3}$. By exploiting the invariance of the length $p_3^2 = p_{3,1}^2 + p_{3,2}^2 + p_{3,3}^2$ this term becomes

$$\delta p_{3,3} = -\frac{(p_{3,1}\delta p_{3,1} + p_{3,2}\delta p_{3,2})}{p_{3,3}} = -\frac{(p_{3,2}\tilde{A} + p_{3,1}p_{2,1})(\hat{e}_1\cdot\delta\vec{\Delta}') + p_{3,2}\tilde{B}(\hat{e}_2\cdot\delta\vec{\Delta}')}{p_{3,3}} \tag{37}$$

This concludes the derivation of $\delta\vec{p}_3$. Inserting Eq. (30, 36, 37, 25, 28, 29) into Eq. (24) we finally obtain

$$\begin{aligned}\delta\vec{p}_3 = \Big[&\quad p_{2,1}\hat{e}_1\hat{e}_1 - p_{3,2}\hat{e}_1\hat{e}_2 - p_{3,3}\hat{e}_1\hat{e}_3 \\ &+\tilde{A}\hat{e}_2\hat{e}_1 + (B + \tfrac{p_{3,2}p_{4,1}}{p_{4,2}})\hat{e}_2\hat{e}_2 + \tfrac{p_{3,3}p_{4,1}}{p_{4,2}}\hat{e}_2\hat{e}_3 \\ &-\tfrac{p_{3,2}\tilde{A}+p_{2,1}p_{3,1}}{p_{3,3}}\hat{e}_3\hat{e}_1 - \tfrac{p_{3,2}\tilde{B}}{p_{3,3}}\hat{e}_3\hat{e}_2 + \tfrac{p_{3,1}p_{4,2}-p_{3,2}p_{4,1}}{p_{4,2}}\hat{e}_3\hat{e}_3 \quad \Big]\cdot\frac{\delta\vec{\Delta}}{\Delta}. \end{aligned} \tag{38}$$

The sum over outer products of basis vectors in the bracket on the right hand side represents a rank two tensor relating $\frac{\delta\vec{\Delta}}{\Delta}$ to $\delta\vec{p}_3$. Since $\delta\vec{\Delta} = -\delta\vec{r}_1$ we can express this relation as

$$\delta\vec{p}_3 = \Pi_3\cdot\delta\vec{r}_1 = \Pi_3\cdot\left(\sum_{j=1}^{n-6}\left[\hat{n}_j\times(\vec{r}_1 - \vec{r}_{pre}^{(j)})\right]\delta\chi_{pre}^{(j)}\right), \tag{39}$$

where Π_3 is readily obtained from Eq. (38) and $\vec{r}_{pre}^{(j)}$ is the position of the atom associated with the j'th prerotational degree of freedom, i.e. the anchoring point of the rotation related to $\chi_{pre}^{(j)}$. With respect to the basis $\{\hat{e}_i\}$, Π_3 has the matrix representation

$$\Pi_3 = \frac{1}{\Delta}\begin{bmatrix} -p_{2,1} & p_{3,2} & p_{3,3} \\ \frac{p_{2,1}}{p_{4,2}}\left[p_{4,1} - A\right] & -\left[B + \frac{p_{3,2}p_{4,1}}{p_{4,2}}\right] & -\frac{p_{3,3}p_{4,1}}{p_{4,2}} \\ \frac{p_{2,1}}{p_{3,3}}\left[\frac{p_{3,2}}{p_{4,2}}A - C\right] & \frac{p_{3,2}}{p_{3,3}}\left[B + C\right] & C \end{bmatrix}, \tag{40}$$

where $C = p_{3,2}\frac{p_{4,1}}{p_{4,2}} - p_{3,1}$ and A and B are defined by Eq. (35). Eq. (39) and Eq. (40) conclude the calculation of the matrix elements $\vec{\Lambda}_{3,j}$. Finally, defining $\Pi_2 = -(\Pi_3 + \mathbb{I})$ as in Eq. (23)), the variation of $\vec{p}_2$ is found as

$$\delta\vec{p}_2 = \Pi_2\cdot\left(\sum_{j=1}^{n-6}\left[\hat{n}_j\times(\vec{r}_1 - \vec{r}_{pre}^{(j)})\right]\delta\chi_{pre}^{(j)}\right) \tag{41}$$

which concludes the calculation of the remaining matrix elements $\vec{\Lambda}_{3,j}$. The full results are summarized in the next section.

Text S3 - M Matrix

The calculations presented above allow us to express the variation of the post-rotational degrees of freedom as $\delta\bar{\chi}_{post} = \mathbf{S}(\bar{\chi}_{pre}, \bar{\chi}_{post})\delta\bar{\chi}_{pre}$, with $S_{k,j} = \sum_{i=0}^{3} \frac{\partial \chi_{post}^{(k)}}{\partial \vec{p}_i} \cdot \frac{\partial \vec{p}_i}{\partial \chi_{pre}^{(j)}} = \sum_{i=0}^{3} \vec{Z}_{k,i} \cdot \vec{\Lambda}_{i,j}$.
$\vec{\mathbf{Z}}$ is a 6×4 matrix whose vectorial elements are given by

$$\vec{\mathbf{Z}} = \begin{pmatrix} \vec{\mathcal{D}}_-(\vec{p}_0,\vec{p}_1,\vec{p}_2,\chi_1) & \vec{\mathcal{D}}(\vec{p}_0,\vec{p}_1,\vec{p}_2,\chi_1) & \vec{\mathcal{D}}_+(\vec{p}_0,\vec{p}_1,\vec{p}_2,\chi_1) & \vec{0} \\ \vec{0} & \vec{\mathcal{B}}_-(\vec{p}_1,\vec{p}_2,\chi_2) & \vec{\mathcal{B}}(\vec{p}_1,\vec{p}_2,\chi_2) & \vec{0} \\ \vec{0} & \vec{\mathcal{D}}_-(\vec{p}_1,\vec{p}_2,\vec{p}_3,\chi_3) & \vec{\mathcal{D}}(\vec{p}_1,\vec{p}_2,\vec{p}_3,\chi_3) & \vec{\mathcal{D}}_+(\vec{p}_1,\vec{p}_2,\vec{p}_3,\chi_3) \\ \vec{0} & \vec{0} & \vec{\mathcal{B}}_-(\vec{p}_2,\vec{p}_3,\chi_4) & \vec{\mathcal{B}}(\vec{p}_2,\vec{p}_3,\chi_4) \\ \vec{0} & \vec{0} & \vec{0} & \vec{\mathcal{B}}_-(\vec{p}_3,\vec{p}_4,\chi_5) \\ \vec{0} & \vec{0} & \vec{0} & \vec{\mathcal{D}}_-(\vec{p}_3,\vec{p}_4,\vec{p}_5,\chi_6) \end{pmatrix}$$

Here, χ_j refers to the j'th postrotational degree of freedom and

$$\vec{\mathcal{B}}_-(\vec{p}_-,\vec{p},\chi) = -\frac{1}{\sin(\chi)}\frac{\vec{p}}{p_- p} \quad ; \quad \vec{\mathcal{B}}(\vec{p}_-,\vec{p},\chi) = -\frac{1}{\sin(\chi)}\frac{\vec{p}_-}{p_- p}.$$

For $\vec{\mathcal{D}}_{-,\cdot,+}$, two sets of equations are available. If $\chi_d \in [\frac{\pi}{4}, \frac{3}{4}\pi[\cup [\frac{5}{4}\pi, \frac{7}{4}\pi[$

$$\begin{aligned} \vec{\mathcal{D}}_-(\vec{p}_-,\vec{p},\vec{p}_+,\chi) &= k_1\left[(\vec{p}\times\vec{b}_-)\cot(\chi) - \frac{\vec{p}\times\vec{b}_+}{\sin(\chi)}\right] \quad ; \quad \vec{\mathcal{D}}_+(\vec{p}_-,\vec{p},\vec{p}_+,\chi) = k_2\left[(\vec{b}_+\times\vec{p})\cot(\chi) + \frac{\vec{p}\times\vec{b}_-}{\sin(\chi)}\right] \\ \vec{\mathcal{D}}(\vec{p}_-,\vec{p},\vec{p}_+,\chi) &= \cot(\chi)\left[k_1(\vec{b}_-\times\vec{p}_-) - k_2(\vec{b}_+\times\vec{p}_+)\right] + \frac{k_1(\vec{p}_-\times\vec{b}_+) - k_2(\vec{p}_+\times\vec{b}_-)}{\sin(\chi)} \end{aligned}$$

otherwise when $\chi_d \in [\frac{3}{4}\pi, \frac{5}{4}\pi[\cup [-\frac{1}{4}\pi, \frac{1}{4}\pi[$

$$\begin{aligned} \vec{\mathcal{D}}_-(\vec{p}_-,\vec{p},\vec{p}_+,\chi) &= k_1\left[p\frac{\vec{b}_+}{\cos(\chi)} - \tan(\chi)(\vec{p}\times\vec{b}_-)\right] \quad ; \quad \vec{\mathcal{D}}_+(\vec{p}_-,\vec{p},\vec{p}_+,\chi) = k_2\left[p\frac{\vec{b}_-}{\cos(\chi)} + \tan(\chi)(\vec{p}\times\vec{b}_+)\right] \\ \vec{\mathcal{D}}(\vec{p}_-,\vec{p},\vec{p}_+,\chi) &= \frac{1}{p\cos(\chi)}\left[(\vec{b}_-\times\vec{b}_+) + k_1(\vec{b}_+\times\vec{p})\times\vec{p}_- + k_2(\vec{b}_-\times\vec{p})\times\vec{p}_+\right] \\ &+ \tan(\chi)\left[k_1(\vec{p}_-\times\vec{b}_-) - k_2(\vec{p}_+\times\vec{b}_+)\right] \end{aligned}$$

where the following quantities have been defined

$$\vec{b}_- = \frac{\vec{p}_-\times\vec{p}}{|\vec{p}_-\times\vec{p}|} \quad ; \quad \vec{b}_+ = \frac{\vec{p}\times\vec{p}_+}{|\vec{p}\times\vec{p}_+|} \quad ; \quad k_1 = \frac{1}{|\vec{p}_-\times\vec{p}|} \quad ; \quad k_2 = \frac{1}{|\vec{p}\times\vec{p}_+|}.$$

$\vec{\mathbf{\Lambda}}$ is a $4 \times (n-6)$ matrix whose elements are given by the vectors

$$\vec{\mathbf{\Lambda}} = \begin{pmatrix} \hat{n}_{pre}^{(1)}\times\vec{p}_0 & \dots & \hat{n}_{pre}^{(n-7)}\times\vec{p}_0 & 0 \\ \hat{n}_{pre}^{(1)}\times\vec{p}_1 & \dots & & \hat{n}_{pre}^{(n-6)}\times\vec{p}_1 \\ \Pi_2\cdot[\hat{n}_{pre}^{(1)}\times(\vec{r}_1 - \vec{r}_{pre}^{(1)})] & \dots & & \Pi_2\cdot[\hat{n}_{pre}^{(n-6)}\times(\vec{r}_1 - \vec{r}_{pre}^{(n-6)})] \\ \Pi_3\cdot[\hat{n}_{pre}^{(1)}\times(\vec{r}_1 - \vec{r}_{pre}^{(1)})] & \dots & & \Pi_3\cdot[\hat{n}_{pre}^{(n-6)}\times(\vec{r}_1 - \vec{r}_{pre}^{(n-6)})] \end{pmatrix}$$

with

$$\hat{n}_{pre}^{(i)} = \begin{cases} \frac{\vec{p}_-^{(i)}\times\vec{p}^{(i)}}{|\vec{p}_-^{(i)}\times\vec{p}^{(i)}|} & \text{if } \chi_{pre}^{(i)} \text{ is bond angle} \\ \frac{\vec{p}^{(i)}}{p^{(i)}} & \text{if } \chi_{pre}^{(i)} \text{ is dihedral angle} \end{cases}$$

Π_2 and Π_3 are given by

$$\Pi_3 = \frac{1}{\Delta}\begin{bmatrix} -p_{2,1} & p_{3,2} & p_{3,3} \\ \frac{p_{2,1}}{p_{4,2}}\left[p_{4,1} - A\right] & -\left[B + \frac{p_{3,2}p_{4,1}}{p_{4,2}}\right] & -\frac{p_{3,3}p_{4,1}}{p_{4,2}} \\ \frac{p_{2,1}}{p_{3,3}}\left[\frac{p_{3,2}}{p_{4,2}}A - C\right] & \frac{p_{3,2}}{p_{3,3}}\left[B + C\right] & C \end{bmatrix} \quad ; \quad \Pi_2 = -(\Pi_3 + \mathbb{I})$$

having defined

$$\begin{aligned} A &= \frac{(\vec{p}_3 \cdot \vec{p}_4)(p_3^2 p_{4,1} - (\vec{p}_3 \cdot \vec{p}_4)p_{3,1}\sin^2(\omega_p)) - p_3^2 p_4^2 p_{3,1}\cos^2(\omega_p)}{D} \\ B &= \frac{p_3^2((\vec{p}_3 \cdot \vec{p}_4)p_{3,1} - p_3^2 p_{4,1})}{D} \\ C &= p_{3,2}\frac{p_{4,1}}{p_{4,2}} - p_{3,1} \\ D &= p_3^2(\vec{p}_3 \cdot \vec{p}_4)\cos^2(\omega_p) - p_{3,1}p_{4,1}p_3^2 + p_{3,1}^2 \vec{p}_3 \cdot \vec{p}_4 \sin^2(\omega_p) \end{aligned}$$

where $p_{i,j}$ indicates the projection of the bond vector $\vec{p}_i$ on the unit vector $\hat{e}_j$ defined in Eq. (1).

The matrix $\mathbf{S}$ constitutes one of the building blocks of the CRISP proposal distribution. Considering Eq. (1) in the main text, we write

$$\begin{aligned} p(\delta\bar{\chi}) &\propto \exp\left\{-\frac{\lambda}{2}(\delta\bar{\chi})^T \mathbf{C}_n (\delta\bar{\chi})\right\} \\ &= \exp\left\{-\frac{\lambda}{2}(\delta\bar{\chi}_{pre} \| \delta\bar{\chi}_{post})^T (\mathbf{C}_{n-6} \oplus \mathbf{C}_6)(\delta\bar{\chi}_{pre} \| \delta\bar{\chi}_{post})\right\} \\ &= \exp\left\{-\frac{\lambda}{2}(\delta\bar{\chi}_{pre} \| \mathbf{S}\delta\bar{\chi}_{pre})^T (\mathbf{C}_{n-6} \oplus \mathbf{C}_6)(\delta\bar{\chi}_{pre} \| \mathbf{S}\delta\bar{\chi}_{pre})\right\} \end{aligned} \tag{42}$$

where $\oplus$ is the matrix direct sum and $\|$ indicates the concatenation of column or row vectors [4]. Therefore

$$\begin{aligned} p(\delta\bar{\chi}_{pre}) &\propto \exp\left\{-\frac{\lambda}{2}(\delta\bar{\chi}_{pre}^T(\mathbf{C}_{n-6} + \mathbf{S}^T\mathbf{C}_6\mathbf{S})\delta\bar{\chi}_{pre})\right\} \\ &= \exp\left\{-\frac{1}{2}(\delta\bar{\chi}_{pre}^T \mathbf{M} \delta\bar{\chi}_{pre})\right\} \end{aligned} \tag{43}$$

which gives the final expression of the proposal distribution of Eq. (3).

[4] Given two row vectors $\bar{a} = [a_1, a_2, \ldots, a_n]$ and $\bar{b} = [b_1, b_2, \ldots, b_m]$ their concatenation is $\bar{a}\|\bar{b} = [a_1, a_2, \ldots, a_n, b_1, \ldots, b_m]$

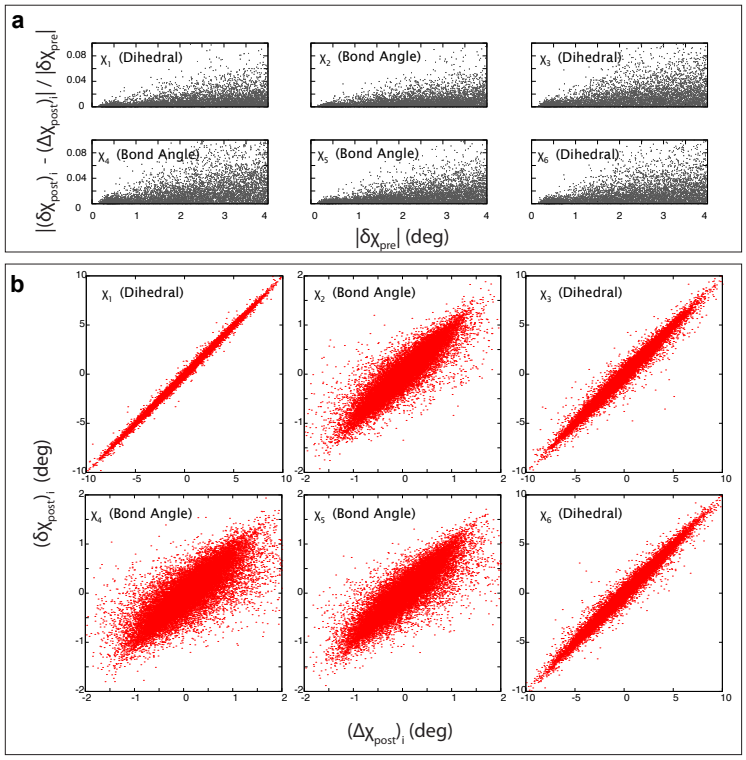


Fig. S1 The first order approximation $\delta\bar{\chi}_{post} = \mathbf{S}\delta\bar{\chi}_{pre}$ is validated by checking that

$$\lim_{|\delta\bar{\chi}_{pre}|\to 0} \frac{|(\delta\chi_{post})_i - (\Delta\chi_{post})_i|}{|\delta\bar{\chi}_{pre}|} = 0 \tag{1}$$

where $\delta\bar{\chi}_{post}$ is the angular variation calculated via the first order approximation and $\Delta\bar{\chi}_{post}$ is the actual change introduced during post-rotation. In the plot we show the quantity defined in Eq.(1) for each degree of freedom, calculated for 25000 CRISP moves on protein G (PDB entry 2GB1). As shown in panel (**a**), the limit in Eq.(1) is satisfied for all the 6 post-rotational degrees of freedom. (**b**) Scatter plot of the actual change $\Delta\bar{\chi}_{post}$ vs. the predicted variation $\delta\bar{\chi}_{post} = \mathbf{S}\delta\bar{\chi}_{pre}$ for the optimal choice of the parameters ($\lambda = 200$ and $k = 65$, Fig. S5). The plots refer to 25000 attempted moves on protein G. With these settings, the average pre-rotation is $< |\delta\bar{\chi}_{pre}| >= 6.13°$, with an average absolute error $< |\delta\bar{\chi}_{post} - \Delta\bar{\chi}_{post}| >= 0.58°$

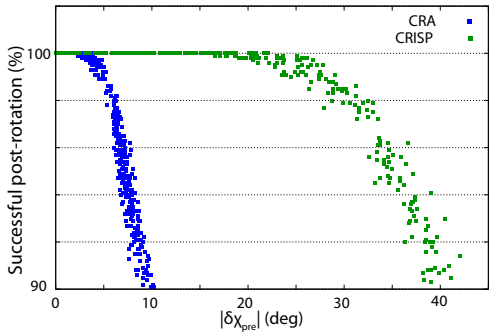


Fig. S2 By tuning the free parameter λ, the allowed angular variation during pre-rotation can be varied both for the CRA and the CRISP move. If the pre-rotation is made too large, however, no solution for chain-closure is available under the constraint of fixed bond length. The selection probability used in CRISP allows to propose tentative updates over a much larger range compared to CRA, without affecting the success of the chain closure. Each point reports the percentage of moves for which a solution for post-rotation is found averaged over 1000 attempted moves on protein G (PDB entry 2GB1).

Text S4 - Outline of the Monte Carlo algorithm

As demonstrated by Theodorou and coworkers,[1] the solution of the concerted rotation problem entails a temporary change in the variables used to describe the geometrical configuration from $\bar{\chi}_{post}$ to the six degrees of freedom of the constraint. This transformation is not metric preserving, and a Jacobian determinant $J(\bar{\chi}_{post})$ relating the volume elements in the two coordinates frames is required in order to preserve detailed balance. This consideration leads to the Metropolis-Hastings acceptance probability for a move $\bar{\chi} \to \bar{\chi}'$ given by

$$P_a(\bar{\chi} \to \bar{\chi}') = \min\left[1, \frac{P_{eq}(\bar{\chi}', \tilde{\chi}) W(\bar{\chi}' \to \bar{\chi}) J(\bar{\chi}'_{post})}{P_{eq}(\bar{\chi}, \tilde{\chi}) W(\bar{\chi} \to \bar{\chi}') J(\bar{\chi}_{post})}\right] \tag{2}$$

where $P_{eq}(\bar{\chi}, \tilde{\chi}) = Z^{-1} \exp(-\beta E(\bar{\chi}, \tilde{\chi}))$ is the Boltzmann equilibrium distribution, E is the energy, and $\tilde{\chi}$ are the degrees of freedom unaffected by the move. $W(\bar{\chi}' \to \bar{\chi})$ is the probability of proposing state $\bar{\chi}'$ when currently in state $\bar{\chi}$.
In this context, the natural choice for the selection probability W is given by Eq. 43. The procedure for a complete CRISP move on n degrees of freedom in a protein of length N is:

- Select a start index s in the range $[1, N - m]$, where m is the number of residues involved in the move.
- Calculate the matrix M for the current structure $\bar{\chi}$.
- Draw a set of variations $\delta\bar{\chi}_{pre}$ from the distribution of Eq. 43 as previously described[2]: the M matrix is first separated by a Cholesky decomposition $M = LL^T$. A vector of independent random numbers $\bar{\psi}$ is then draw from a Gaussian distribution with zero mean and unit variance. Finally, the system $L^T \delta\bar{\chi}_{pre} = \bar{\psi}$ is solved for $\delta\bar{\chi}_{pre}$.
- Modify the pre-rotational degrees of freedom $\bar{\chi}'_{pre} = \bar{\chi}_{pre} + \delta\bar{\chi}_{pre}$.
- Find the solution for the chain closure using the analytical solution, leading to a new conformation $\bar{\chi}' = (\bar{\chi}'_{pre}, \bar{\chi}'_{post})$.

- If a solution is found, calculate the ratio $\frac{J(\bar{\chi}'_{post})W(\bar{\chi}'\rightarrow\bar{\chi})}{J(\bar{\chi}_{post})W(\bar{\chi}\rightarrow\bar{\chi}')}$.
- Accept the move with the probability given by Eq. (2).

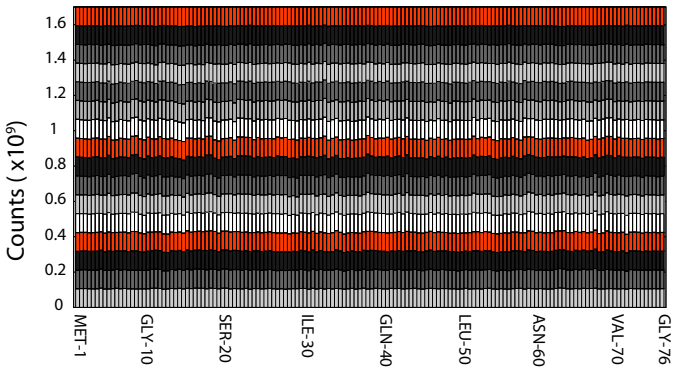


Fig. S3 We verify the correct implementation of CRISP move by checking that the dihedral degrees of freedom are uniformly distributed in the absence of a force field. The histogram shows the dihedral angle population for all the ϕ, ψ angles of protein ubiquitin in the absence of the force field using CRISP moves. Each bar represents the counts in one of the 16 bins for each dihedral angle in the chain.

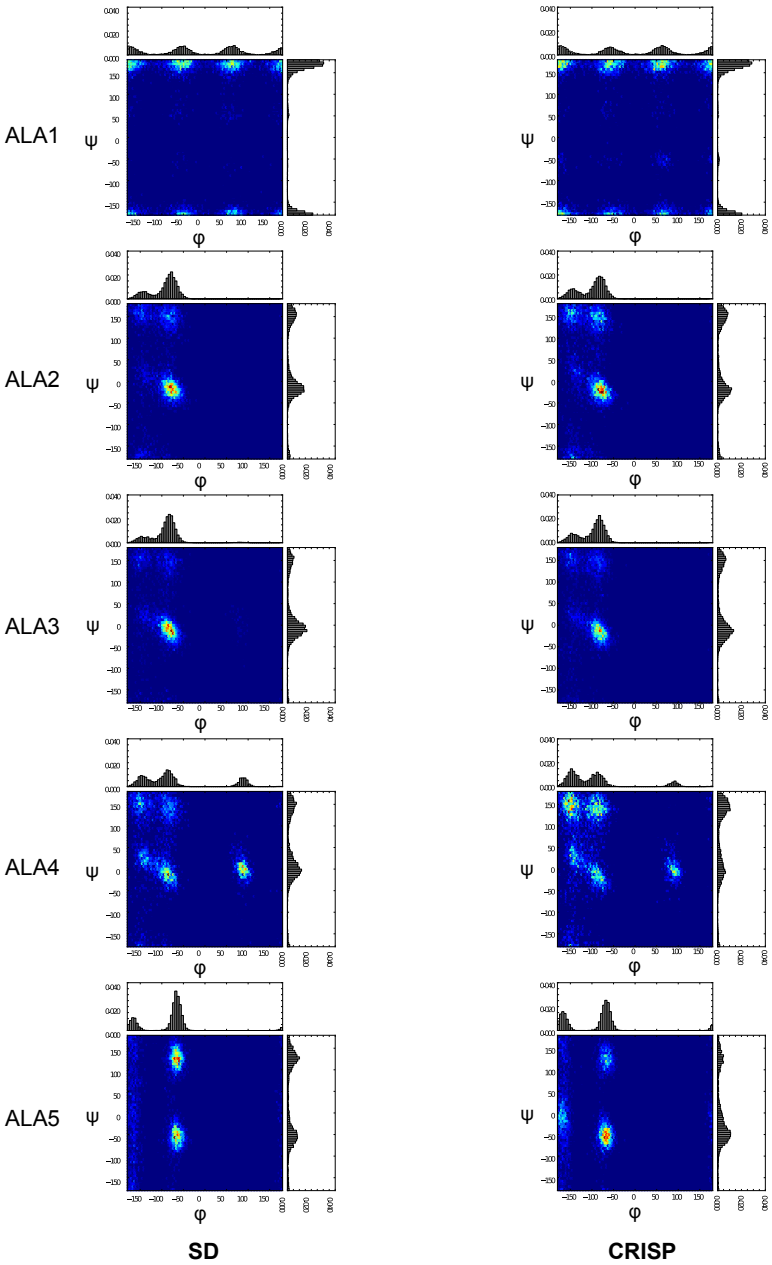


Fig. S4 We verify the correctness of the MC approach by comparing the angular distribution obtained from $5 \times 100ns$ SD simulations performed with TINKER on Alanine 5 (left column) with the angular distribution obtained from 5×10^8 steps MC simulations using CRISP (right column). The same test was performed using CRA and CRANKSHAFT, giving similar results.

Table S1 Comparison between average collective variables calculated using SD and different MC methodologies. Averages and standard error of the average calculated on 5×10^8 MC steps / $5 \times 100ns$ simulations on Ala_5.

	Energy (Kcal/mol)*	Radius of Gyration (Å)	RMSD (Å)†
SD	-258.367 ± 1.113	3.754 ± 0.148	1.787 ± 0.099
CRA	-256.856 ± 0.824	3.820 ± 0.117	1.809 ± 0.106
CRANKSHAFT	-253.969 ± 0.572	3.866 ± 0.072	1.764 ± 0.081
CRISP	-255.847 ± 0.498	3.857 ± 0.066	1.809 ± 0.029

* OPLSaa potential + GB/SA solvation energy.

† The alignment is performed on Ala_5 in helical conformation

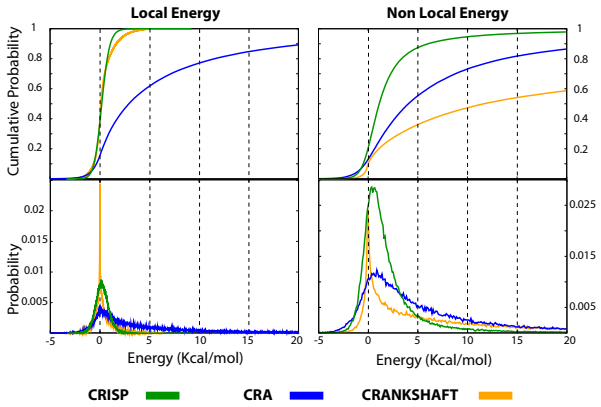


Fig. S5 Probability distribution and cumulative probability distribution of energy jumps based on 10^5 proposed CRA, CRISP and CRANKSHAFT local moves on protein ubiquitin, using the OPLS_{aa} potential in combination with the GB/SA solvation model. For the local part of the energy (angle bend and dihedral energy), all the distributions are characterized by one mode around zero. CRA shows a heavy right-side tail, as a consequence of the large variations in bond angles introduced by this move during post-rotation. CRISP and CRANKSHAFT behave very similarly, however it must be highlighted that in each single CRISP move 12 bond and 12 dihedral angles are modified, while in CRANKSHAFT moves 4 dihedral and 2 bond angles are varied at a time. Non local interactions (charge-charge, van der Waals and solvation energy) contributes to a similar extent to the total energy jump. As shown in the plots on the right, concerted-rotation methods introduce less dramatic variations compared to CRANKSHAFT move, for which, in the 50% of the cases, a variation in energy larger than 10 Kcal/mol is introduced, corresponding to an acceptance probability $< 0.25\%$ at physiological temperature.

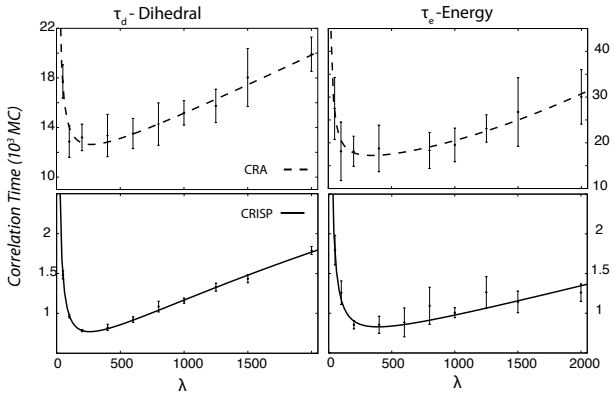


Fig. S6 Dependence of the average correlation time $\bar{\tau}_d$ of the 20 central dihedral angles of Ala_{14} and of the energy, τ_e, on the choice of the parameter λ (In the notation introduced in the original CRA paper the parameter λ corresponds to c_1. The values of the other CRA parameters were set to $c_2 = 8$ and $c_3 = 20$ as described in the original paper). The solid line is a polynomial fit of the data to guide the eye. For each value of λ, 5 independent runs were started in the α-helical conformation, fluctuating around that state throughout the simulation. Correlation times were calculated using the Jackknife analysis[3] by subdividing the N observations of the quantity of interest, o, in N_B blocks of length l containing all data except one of the previous binning blocks $o_{J,n} = \frac{N\bar{o} - \sum_{i=1}^{l} o_{(n-1)l+i}}{N-l}$. In the Jackknife analysis, the correlation time τ of the observable o is estimated as $\tau = \lim_{l\to\infty} \frac{N(N_B-1)^3}{N_B} \frac{\sigma_J^2}{2\sigma_o^2} = \lim_{l\to\infty} f(l)$, where σ_J^2 is the block variance and σ_o^2 the sample variance. For $N \gg l$, we observed that $f(l)$ is a monotonically increasing function of l. We therefore estimated τ by fixing $N = 10^9$ and ensuring the convergence of f as $l \to \infty$. For both CRISP and CRA the value of the parameter λ that minimize the correlation time is $\lambda_{opt} = 200$, which was used in all simulations carried out in this work. The optimal choice of the auxiliary parameter k, which set the scale the allowed variations of bond and peptide dihedral angle with respect to the variations of ϕ and ψ (c.f. Eq. (2)) in the main text, was determined by choosing k such that the average dihedral stepsize $\langle \delta\omega \rangle = \frac{1}{N} \sum_{i=1}^{N} |\delta\omega_i|$ is maximized, where $|\delta\omega_i| = \sqrt{\sum_{j=1}^{n_\omega} \delta\omega_j}$ is the average dihedral variation per local move. This procedure led to an optimal $k_{opt} = 65$.

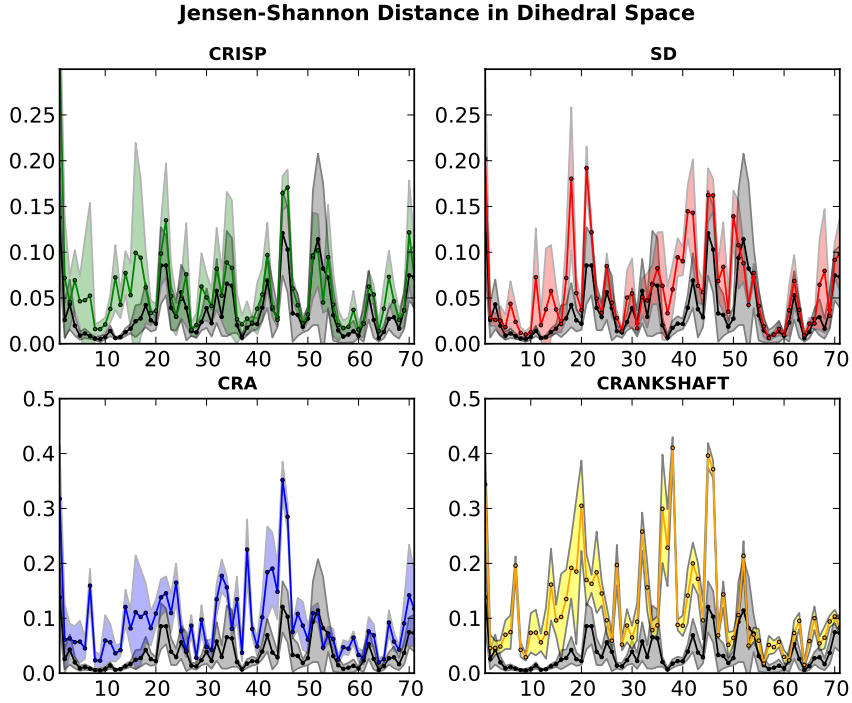


Fig. S7 Average Jensen-Shannon divergence in ϕ/ψ space between the ensemble composed by the all the MD structures ($\overline{MD}$) and SD, CRISP, CRA, CRANKSHAFT simulations. The average distance between single MD simulations and the $\overline{MD}$ ensemble is shown in black/gray. Averages and standard deviations (illustrated as shaded regions) are calculated over ten independent runs. This detailed residue-wise measure of ensemble differences provides additional support for the conclusion in the main text that the CRISP method more accurately reproduces the MD ensemble than existing MC-methods.

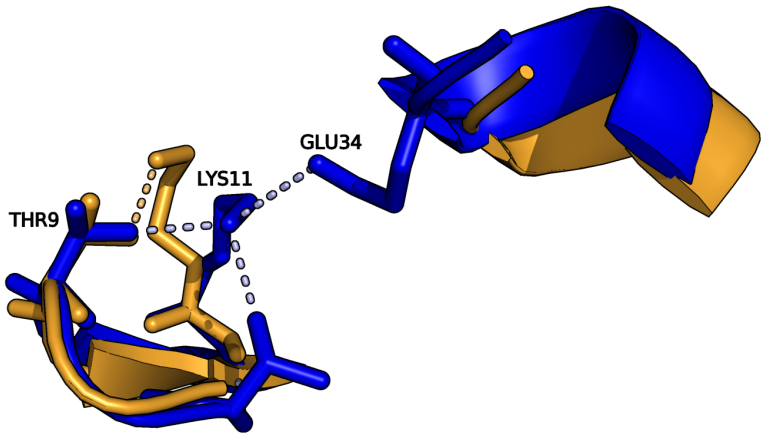


Fig. S8 Snapshots from a Monte Carlo simulation on ubiquitin, illustrating the sidechain sampling problem. Throughout the simulation, LYS11 visits different rotamer states such as the one shown in yellow, in which only one hydrogen bond with THR9 is present, and in blue, where three non-covalent interactions are formed. In this situation any sudden rotameric jump from the latter state would be energetically unfavorable, and as an effect the sidechain is locked in this state. Inspired by the Biased Gaussian Step method[2], we overcome this difficulty by introducing sidechain moves constrained towards small displacement of atoms involved as acceptors or donors in hydrogen bonds This type of move was found to greatly facilitate the inter-conversion between different rotamer states, especially for buried residues involved in the dense network of non-covalent interactions.

References

(1) Dodd, L.; Boone, T.; Theodorou, D. *Mol. Phys.* **1993**, *78*, 961-996.
(2) Favrin, G.; Irback, A; Sjunnesson, F. *J. Chem. Phys.* **2001**, *114*, 8154-8158.
(3) Janke, W. Statistical Analysis of Simulations: Data Correlations and Error Estimation.
In *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, John von Neumann Institute for Computing: Jülich, Germany, 2002.

2.2 PHAISTOS: A Framework for Markov Chain Monte Carlo Simulation of Proteins

Although based on a very simple idea, the development, design and implementation of a Monte Carlo software package for biomolecular simulations entails a number of theoretical and technical challenges. Together with Dr. Wouter Boomsma (Department of Astronomy and Theoretical Physics, University of Lund, Sweden) and the group of Structural Bioinformatics led by Prof. Thomas Hamelryck (Department of Biology, University of Copenhagen, Denmark) we developed PHAISTOS, a framework for Markov chain Monte Carlo simulation of proteins. In this research article we present the main features of the package and we discuss current applications of the software.
This is a co-author article, I was involved in the design and development of the code used to perform many of the trial MC updates, and I implemented parts of the $OPLS_{AA}$ force-field and GB/SA solvent model. Moreover, I did extensive debugging and testing of the package, and contributed to the code documentation.

PHAISTOS: A Framework for Markov Chain Monte Carlo Simulation of Proteins

Wouter Boomsma*†‡, Jes Frellsen§, Tim Harder§, Sandro Bottaro‡, Kristoffer E. Johansson§, Pengfei Tian‡, Kasper Stovgaard§, Christian Andreetta§, Anders S. Christensen¶, Simon Olsson§, Jan Valentin§, Mikael Borg§, Jesper Ferkinghoff-Borg‡, Thomas Hamelryck§

March 27, 2012

Abstract

We present a new framework for conducting Markov chain Monte Carlo sampling for protein simulation, prediction, and structural inference from experimental data. The software package contains implementations of a number of recent advances in Monte Carlo methodology, such as highly efficient local kinetics and sampling from probabilistic models of local protein structure. Currently, two established force-fields are available within the framework, the PROFASI effective forcefield, and OPLS$_{AA}$ with the GB/SA implicit solvent model. A flexible command-line and configuration file interface allows users to quickly set up a simulation with the desired settings.

PHAISTOS is released under the GNU General Public License v3.0. Source code and documentation are freely available at `http://phaistos.sourceforge.net`. The software is implemented in C++, and has been tested on Linux and OS X platforms.

Keywords: Protein, simulation, software, Markov chain Monte Carlo, PHAISTOS,

*to whom correspondence should be addressed
†Department of Astronomy & Theoretical Physics, University of Lund, SE-223 62 Lund, Sweden
‡Department of Biomedical Engineering, DTU Elektro, DTU, 2800 Kgs. Lyngby, Denmark
§Department of Biology, University of Copenhagen, 2200 Copenhagen N, Denmark
¶Department of Chemistry, University of Copenhagen, 2100 Copenhagen E, Denmark

PHAISTOS

We present a new software package for conducting protein simulations. The PHAISTOS framework contains implementations of a range of novel sampling techniques developed over the last years, which are now made available for the first time. The package provides tools for a variety of different tasks, including reversible folding simulations and prediction/determination of structure from experimental data. The code is released under an open source license and full documentation is available online.

INTRODUCTION

Two methods dominate the field of molecular simulation: molecular dynamics (MD) and Markov chain Monte Carlo (MCMC). MD involves iteratively calculating the forces exerted on each particle in a system, and using Newton's equations of motion to update their positions. In contrast, MCMC is a statistical approach, where the goal is to generate samples from the Boltzmann distribution associated with the system. MD has typically been regarded as best-suited for exploring the native ensemble of dense systems, while MCMC methods are typically used for longer time scale simulations. However, using improved move sets, it has recently been demonstrated that even in the densely packed native state, MCMC can serve as an efficient alternative to MD[1].

Although several publicly available MCMC simulation packages exist[2–5], these packages have not generally obtained the same broad acceptance in the scientific community as some of the best-known MD packages. The framework presented in this paper contains implementations of various recently developed tools that increase the efficiency of MCMC-based simulations. By making our methods available in an easily extendible framework, we hope to further encourage the use of MCMC for protein simulations, and promote the development of new MCMC methodologies for the simulation, prediction and inference of protein structure.

METHODOLOGY

The main distinguishing feature of the PHAISTOS package is efficient sampling, obtained through an elaborate set of both established and novel Monte Carlo moves. Complemented with an interface to the MUNINN generalized ensemble MCMC program (`http://muninn.sourceforge.net/`), this constitutes an ideal tool for rapidly exploring protein conformational space. The set of available moves include various types of pivot moves, the crankshaft/backrub local move[6,7], the CRA local move[8], and the semi-local biased Gaussian move[9]. Side-chain sampling can be done either from Gaussian distributions given by rotamer libraries[10] or by sampling values locally around the current state. Finally, PHAISTOS contains a number of unique new moves, described in the sections below. Note that all moves

in PHAISTOS can be applied so that they obey detailed balance, which ensures, together with the ergodicity provided by MUNINN-based sampling, that simulations sample from a well-defined target distribution.

The user can specify all simulation settings directly from the command line or in a configuration file, making it possible to quickly set up a simulation. In particular, this allows the user to readily experiment with different force-fields and parameter values. Currently, two well-known force-fields are available: PROFASI[4], and the OPLS$_{AA}$[11] forcefield in combination with the GB/SA implicit solvent model[12].

PHAISTOS also serves as an easily extendible library for the development of tools for protein structure simulation, inference and prediction. One of the design goals in PHAISTOS is to ensure that new energy terms and moves can be easily implemented with little knowledge of the overall code. Iterators are provided for easy iteration over atoms in a molecule. In addition, caching and rapid determination of interacting atom pairs is made possible by an implementation of the chaintree algorithm[13]. Through a modular build-system, users can readily write their own modules utilizing the library.

Probabilistic models

We recently developed a number of probabilistic models that describe protein structure on a local length scale. A unique feature of the models is that they allow sampling of protein-like conformations in continuous space, providing a statistically rigorous alternative to the widely used fragment and rotamer libraries. Three different models are available: FB5HMM covers the Cα trace of a protein[14], while TORUSDBN and BASILISK, respectively, model backbone and side-chain structure in atomic detail[15,16]. All models can be applied both as proposal distributions (moves) and as components of an energy function.

Efficient local kinetics

To ensure efficient sampling in densely packed environments, PHAISTOS includes the CRISP method, a novel Monte Carlo move that produces local deformations in a small segment of a protein chain. Unlike other local move approaches[7,8], CRISP makes it possible to generate

small updates to the chain without disrupting its local geometry. We have recently shown that this can have a dramatic impact on simulation performance, in particular around the densely packed native state[1].

Generalized Ensembles

In addition to fixed temperature Metropolis-Hastings (MH) simulations, PHAISTOS includes support for conducting simulations in generalized ensembles. While MH simulations generate samples from the Boltzmann distribution, generalized ensemble methods can avoid convergence and ergodicity problems by sampling from a modified distribution, after which sampled structures can be reweighted to obtain the original Boltzmann distribution at any given temperature. We have recently developed an automated method, MUNINN, for estimating weights in generalized ensemble simulations (`http://muninn.sourceforge.net/`). It employs the generalized multi-histogram equations[17], and uses a non-uniform adaptive binning of the energy space, ensuring efficient scaling to large systems. Optionally, weights can be restricted to cover a limited temperature range of interest.

RESULTS

To illustrate the versatility of PHAISTOS, we highlight several recently published applications of the framework.

Reversible folding and native ensembles

The OPLS$_{AA}$[11] and PROFASI[4] energy functions, available in PHAISTOS, represent two extremes in the range of force-fields available in the literature: an ultrafast force-field modeling effective interactions in a solvent, which has been successfully used for reversible folding simulations of a range of proteins[4], and a highly detailed classic molecular mechanics force-field, which is useful in the exploration of details of native ensembles. The energy landscape around the native state tends to be rugged, and it can be challenging to sample such states efficiently. For such tasks, the CRISP move is particularly well suited, given its ability to pro-

pose subtle, non-disruptive updates to the protein backbone. Monte Carlo simulations using this move were recently shown to perform on par with molecular dynamics, outperforming the current state-of-the-art in local move methods[1].

The TYPHON program[18] (implemented as a PHAISTOS module) rapidly explores near-native ensembles by using the CRISP move in combination with a user-defined set of non-local restraints. Local structure is under the control of probabilistic models of the backbone (TORUSDBN) and side chain (BASILISK), while non-local interactions such as hydrogen bonds and disulfide bridges are imposed as Gaussian restraints. Typhon can be seen as a "null model" of conformational fluctuations in proteins: it rapidly explores the conformational space accessible to a protein given a set of specified restraints.

Structure prediction and determination

PHAISTOS can also be applied in the context of protein structure prediction and inference from experimental data. A recent study demonstrated the prediction aspect using a combination of TORUSDBN with probabilistic models of compactness and hydrogen bonding[19].

The framework also includes support for simulations restrained with various types of experimental data. Small angle X-ray scattering (SAXS) is an experimental technique that provides low resolution information on the overall shape of a protein. It is particularly useful for determining the relative orientations in multi-domain proteins or complexes. PHAISTOS contains a SAXS-module that provides support for inferring the structure of proteins consisting of multiple domains connected by flexible linkers, given the atomic structures of the individual domains. The method relies on the efficient back-calculation of SAXS curves based on a coarse grained Debye method[20]. Furthermore, PHAISTOS was recently used for inferential structure determination using NOE data[21].

Efficient clustering

Efficient clustering of large numbers of protein structures is an important task in protein structure prediction and analysis. Typically, clustering programs require the costly calculation of the root mean square deviations (RMSD) for many pairs in the set of structures.

PHAISTOS contains a clustering module called PLEIADES, that uses a K-means clustering approach to avoid the calculation of pairwise RMSD calculations. Furthermore, the RMSD distance computations can be replaced with distances between vectors of Gauss integrals[22], a technique which provides dramatic computational speedups[23].

DISCUSSION

The PHAISTOS framework provides a set of tools for conducting MCMC simulations of protein systems, incorporating efficient conformational sampling and generalized ensembles. The sampling efficiency is ensured through an extensive set of Monte Carlo moves. This includes both a novel local move, and several biased proposal moves, in which the sampling is controlled by probabilistic models of local structure, a technique which is unique to this framework[14–16].

In conclusion, PHAISTOS extends the scope of MCMC methods for protein simulation, prediction and structural inference. The software is freely available, providing the scientific community with a versatile toolkit for a wide variety of *in silico* protein challenges. The source code is fully documented using the Doxygen system, and a user manual is available for detailed descriptions on how simulations are set up. Both sources of information are accessible via the PHAISTOS web site `http://phaistos.sourceforge.net`.

ACKNOWLEDGMENTS

This work was supported by the Danish Council for Independent Research [FNU272-08-0315 to W.B., FTP274-06-0380 to K.S, FTP09-066546 to S.O., J.V., FTP274-08-0124 to K.E.J.], the Danish Council for Strategic Research [NABIIT2106-06-0009 to J.F., Ti.H., C.A., M.B.], the Novo Nordisk STAR Program [A.S.C.], and Radiometer (DTU) [S.B.].

References

1. S. Bottaro, W. Boomsma, K. E. Johansson, C. Andreetta, T. Hamelryck, and J. Ferkinghoff-Borg, J. Chem. Theory Comput. **8**, 695 (2012).

2. M. Martin and J. Siepmann, J. Phys. Chem. B **103**, 4508 (1999).

3. J. Hu, A. Ma, and A. Dinner, J. Comput. Chem. **27**, 203 (2006).

4. A. Irbäck, S. Mitternacht, and S. Mohanty, PMC Biophysics **2**, 2 (2009).

5. A. Vitalis and R. Pappu, Annu. Rep. Comput. Chem. **5**, 49 (2009).

6. M. Betancourt, J. Chem. Phys. **123**, 174905 (2005).

7. C. Smith and T. Kortemme, J. Mol. Biol. **380**, 742 (2008).

8. J. Ulmschneider and W. Jorgensen, J. Chem. Phys. **118**, 4261 (2003).

9. G. Favrin, A. Irbäck, and F. Sjunnesson, J. Chem. Phys. **114**, 8154 (2001).

10. R. L. Dunbrack and F. E. Cohen, Protein Sci. **6**, 1661 (1997).

11. D. S. Jorgensen, W.L. Maxwell and J. Tirado-Rives, J. Am. Chem. Soc. **18**, 11225 (1996).

12. D. Qiu, S. P.S., F. Hollinger, and W. Still, J. Phys. Chem. A **101**, 3005 (1997).

13. I. Lotan, F. Schwarzer, D. Halperin, and J. Latombe, J. Comput. Biol. **11**, 902 (2004).

14. T. Hamelryck, J. Kent, and A. Krogh, Plos. Comput. Biol. **2**, e131 (2006).

15. W. Boomsma, K. Mardia, C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck, Proc. Natl. Acad. Sci. U. S. A. **105**, 8932 (2008).

16. T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. Johansson, and T. Hamelryck, BMC Bioinformatics **11**, 306 (2010).

17. J. Ferkinghoff-Borg, J. Eur. Phys. J. B. **29**, 481 (2002).

18. T. Harder, M. Borg, S. Bottaro, W. Boomsma, S. Olsson, J. Ferkinghoff-Borg, and T. Hamelryck, Structure (2012), in press.

19. T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, and J. Ferkinghoff-Borg, PLoS ONE **5**, e13714 (2010).

20. K. Stovgaard, C. Andreetta, J. Ferkinghoff-Borg, and T. Hamelryck, BMC Bioinformatics **11**, 429 (2010).

21. S. Olsson, W. Boomsma, J. Frellsen, S. Bottaro, T. Harder, J. Ferkinghoff-Borg, and T. Hamelryck, J. Magn. Reson. **213**, 182 (2011).

22. P. Røgen and B. Fain, Proc. Natl. Acad. Sci. U. S. A. **100**, 119 (2003).

23. T. Harder, M. Borg, W. Boomsma, P. Røgen, and T. Hamelryck, Bioinformatics **28**, 510 (2012).

2.3 An Efficient Null Model for Conformational Fluctuations in Proteins

In this research paper we present a simple yet accurate Monte Carlo method to characterize the flexibility of proteins in near-native conditions. The approach is based on the idea of conducting a Monte Carlo simulation using probabilistic models of local structure to describe the ϕ, ψ backbone and χ side-chain angles propensities of proteins, while imposing a set of user-defined restraints (*e.g.* hydrogen bonding and disulfide bridges) to model the non-local interactions of near-native conformations. In this context, the conformational space is efficiently explored with the use of CRISP moves. The method is validated by comparing the native dynamics obtained from the simulations with experimental measurements.

This is a co-author article, I was involved in the design and implementation of the method used to combine the MC moves together with the probabilistic model of local structure. In addition, I contributed to analyze the trajectories obtained from the MC simulations.

Structure

Article

An efficient null model for conformational fluctuations in proteins

Tim Harder[1], Mikael Borg[1], Sandro Bottaro[2], Wouter Boomsma[2,3], Simon Olsson[1], Jesper Ferkinghoff-Borg[2], Thomas Hamelryck[*1]

[1] The Bioinformatics Section, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[2] DTU Elektro, Technical University of Denmark, Lyngby, Denmark
[3] Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden
[*] Correspondence: thamelry@binf.ku.dk

SUMMARY

Protein dynamics plays a crucial role in function, catalytic activity and pathogenesis. Consequently, there is great interest in computational methods that probe the conformational fluctuations of a protein. However, molecular dynamics simulations are computationally costly and therefore often limited to comparatively short timescales.

TYPHON is a novel method to explore the conformational space of proteins under the guidance of a probabilistic model of local structure and a given set of restraints that represent nonlocal interactions such as hydrogen bonds or disulfide bridges. The choice of the restraints themselves is heuristic, but the resulting probabilistic model is well-defined and rigorous. Conceptually, TYPHON constitutes a null model of conformational fluctuations under a given set of restraints.

We demonstrate that TYPHON can provide information on conformational fluctuations that is in correspondence with experimental measurements. TYPHON provides a flexible yet computationally efficient method to explore possible conformational fluctuations in proteins.

HIGHLIGHTS

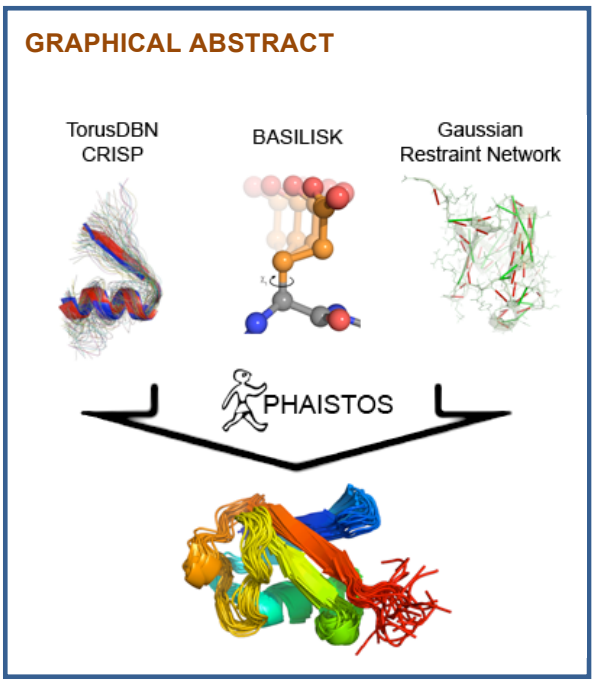
- TYPHON explores a protein's conformational space under given non-local restraints.
- TYPHON's advanced probabilistic models ensure protein-like local structure.
- TYPHON can reproduce a range of experimental results.
- TYPHON's computational efficiency makes large

INTRODUCTION

Over the past few decades it has become increasingly accepted that proteins are dynamic molecules (Stryer 1988). While many proteins adapt unique and specific folds, their inherent flexibility is often essential to the protein's function. However, flexibility can also lead to pathogenesis through misfolding, possibly leading to the formation of aggregates and fibrils (Dobson 2003; Teilum et al. 2009a).

Computer simulations have emerged as important tools to study the dynamics of proteins, complementing the data obtained from biophysical experiments. A variety of methods are available, ranging from detailed, all atom molecular dynamics (MD) simulations (McCammon et al. 1977; Karplus and McCammon 2002; Hess et al. 2008) to coarse grained and approximative methods such as normal mode analysis (NMA) (Levitt et al. 1983), elastic networks (Zheng et al. 2007), tCONCOORD (de Groot et al. 1997; Seeliger and Groot 2009) and FRODA (Jacobs et al. 2001; Wells et al. 2005). All methods come with a trade off between the level of detail and the computational cost for obtaining useful information.

The concept behind MD simulations is to approximate the physical forces acting on a protein and to calculate the motion of particles in the system by applying Newton's laws of motion (McCammon et al. 1977; Karplus and McCammon 2002; Hess et al. 2008). Since the calculation of these physical forces is computationally expensive, MD simulations are usually limited to short timescales — typically in the range of hundreds of nano second. The high level of detail in MD simulations make general physical conclusions viable (van Gunsteren et al. 1996; Brooks et al. 2009). However, the timescales routinely accessible through MD simulations rarely cover the full dynamic range of proteins. Coarse grained MD simulations sacrifice certain atomic details to gain a computational advantage, thus allowing longer simulation times or simulations of larger systems. Merging multiple atoms into so called *beads* or *pseudo atoms* is a common approach to reduce the number of particles in the system (Marrink et al. 2007). Another solution to overcome the computational cost of MD simulations is to use faster computer hardware. Shaw and co-workers were able to achieve a millisecond simulation using custom built, special purpose hardware (Klepeis et al. 2009; Shaw et al. 2010).



Many faster, heuristic alternatives to MD have been developed. The idea behind elastic network (EN) models is that the dynamics of folded, native proteins are rather limited compared to unfolded dynamics, and overall governed by the inter-residue contact topology (Bahar and Rader 2005). Over the past years, the computationally efficient EN models have replaced the original harmonic potentials in many NMA approaches (Bahar and Rader 2005; Yang et al. 2009). In EN models, the protein's atoms are viewed as point masses that are interconnected by springs. Often only the backbone C_α atoms are included. Subsequently, a number of conformations are sampled and a principal component analysis is performed on the generated ensemble, yielding the normal modes (Levitt et al. 1983). However, ensembles sampled from EN models can also be used in different scenarios (Zheng et al. 2007); *vice versa*, normal modes can also be calculated from ensembles generated in MD simulations (Hess et al. 2008).

Other heuristic approaches that include atomic detail have gained popularity over the past years. FRODA (Jacobs et al. 2001; Wells et al. 2005) identifies rigid substructures in the protein structure to reduce the degrees of freedom for the subsequent simulation. Another widely used, heuristic tool is tCONCOORD (de Groot et al. 1997; Seeliger et al. 2007; Seeliger and Groot 2009), which has successfully been applied in different contexts (Zachariae et al. 2008; Seeliger and De Groot 2010). Here, the input structure is analyzed to create a network of constraints. Subsequently, tCONCOORD randomly perturbs the atom coordinates within a box around their initial positions in the native structure. Then, a Monte Carlo procedure changes the perturbed atomic positions until they again satisfy the constraints. In this procedure, the atomic positions are subject to changes sampled from a uniform distribution. Consequently, all the information is encoded in the constraint network; in the absence of constraints there is no information on how to arrange the atoms.

Here, we present TYPHON, which adopts a probabilistic approach to exploring conformational fluctuations in proteins. TYPHON is based on two recent innovations: TorusDBN (Boomsma et al. 2008) and BASILISK (Harder et al. 2010). TorusDBN and BASILISK are probabilistic models of the conformational space of a protein's main chain and its amino acid side chains, respectively. Both models are formulated as dynamic Bayesian networks (DBNs), and make use of directional statistics (Mardia and Jupp 2000) — the statistics of angles and directions — to represent protein structure in a natural, continuous space (Hamelryck et al. 2006; Boomsma et al. 2008; Harder et al. 2010). Together, TorusDBN and BASILISK constitute a probabilistic model of protein structure in atomic detail. This model is *generative*; plausible protein conformations can be efficiently sampled. Furthermore, TYPHON incorporates CRISP (Bottaro et al. 2012), an efficient method for applying local modifications to the protein's conformation.

The application of these probabilistic models in TYPHON ensures that the structure remains protein-like on a local length scale throughout the conformational sampling. The long-range structure is maintained by imposing different types of distance based restraints, which are heuristic representations of nonlocal interactions such as hydrogen bonds. TYPHON uses Gaussian distributions to implement the restraints, resulting in a valid probabilistic description of the restraint network and the local structure of proteins. This well justified probabilistic formulation differs from previous *ad hoc* approaches. TYPHON explores the conformational space accessible to a protein, within the limits imposed by the restraint network. In the absence of a restraint network, sampling is solely guided by the probabilistic models and results in an ensemble of extended conformations with realistic local structure, conceptually reminiscent of an "unfolded state".

In short, TYPHON can be considered a null model of conformational fluctuations given a set of probabilistic restraints. We again stress that our method is well justified *given* a chosen set of restraints; the biological relevance of the obtained conformations will necessarily depend on the relevance of the heuristic restraints. However, TYPHON provides default restraints, which typically deliver good results for common applications, as discussed below.

In the following, we compare results obtained from TYPHON with experimental measures describing the native ensemble of folded proteins, including B-factors, nuclear magnetic resonance (NMR) order parameters and residual dipolar couplings (RDCs). The different measures allow us to investigate how well TYPHON captures the flexibility of a folded protein. We then demonstrate how local unfolding caused by the loss of metal ions is correctly modeled by TYPHON. Finally, we show how fluctuations of local structure can be investigated under the control of the probabilistic models, which is an additional attractive and innovative aspect of our approach.

RESULTS

Overview of TYPHON

TYPHON samples protein structures from a joint probability distribution that includes local and nonlocal interactions (described in more detail in Experimental procedures). TYPHON incorporates several sophisticated probabilistic models to maintain the local structure, and uses simple Gaussian restraints to maintain relevant non-local interactions. Although the choice of these non-local restraints is heuristic, the resulting joint probabilistic model is well defined and rigorous. In other words, if a suitable restraint network can be chosen for the problem of interest, TYPHON will typically deliver good results, obtained from a well defined probability distribution.

By default, TYPHON automatically detects the hydrogen bond network. The geometry of the individual hydrogen bonds is restrained using a simple model based on four distances modeled by Gaussian probability distributions. Disulfide bridges are by default treated in a similar way. By default, TYPHON also restrains all distances between C_α atoms that are five or more residues apart in the amino acid chain, and within six Å of each other. The latter restraints aim to capture general interactions that stabilize the protein, such as the hydrophobic effect.

The user can manipulate and verify the restraint network. For example, it is possible to disregard all hydrogen bonds involving side chains, or to add or remove restraints between arbitrary atom pairs. In this manuscript, we use different restraint networks to answer different questions. These networks range from involving C_α atoms (see Experimental B-factors) over hydrogen bonds (see Generating a native ensemble) to a small number of disulfide bridges (see Local structure under the control of probabilistic models).

TYPHON is obviously limited with respect to modeling the formation and dissolution of non-local interactions themselves, as the restraint network is fixed throughout the sampling procedure. However, the secondary structure can to some extent be put under the control of the probabilistic models (see Local structure under the control of probabilistic models), allowing for formation and dissolution of certain hydrogen bonds, notably in helices.

Experimental B-factors

The Protein Data Bank (PDB) (Berman et al. 2000) currently contains over 77,000 solved structures; the majority of them are determined by X-ray crystallography. Experimental B-factors associated with the atoms of a crystal structure often give a first indication of the conformational fluctuations within a protein. The B-factor reflects both the thermal vibrations of single atoms and small structural differences between molecules in the crystal. The latter contribution is of interest for inferring protein flexibility. In this test, we analyze whether TYPHON is able to reproduce the flexibility that is indicated by the B-factors of a protein.

TYPHON makes it possible to sample an ensemble of structures that is close to the native structure. We illustrate this with the crystal structure of the 317 residue long protein *Candida antarctica* Lipase B (CalB, PDB: 1tca) (Uppenberg et al. 1995). CalB is an enzyme with industrial applications that adopts an α/β fold. A short helix, consisting of residues 139 to 147, is suspected to act as a flexible lid that is important for catalysis, making it a prime subject of dynamics studies (Skøt et al. 2009). For comparison, we translated the experimental B-factors of the crystal structure into root-mean-square fluctuations (RMSF) using the following relation (Kuzmanic and Zagrovic 2010):

$$RMSF_i^2 = \frac{3B_i}{8\pi^2}$$

where B_i is the B-factor for the *i*-th residue.

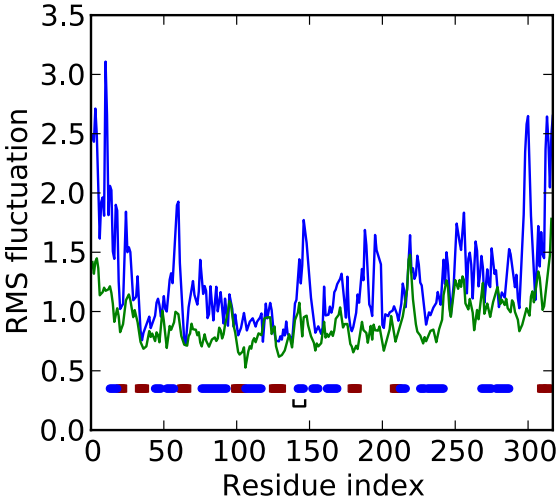


Figure 1: Experimental B-factors of Candida antarctica lipase B. The figure shows root-mean-square fluctuations calculated from the B-factors taken from the crystal structure (PDB: 1tca, green line) and calculated from a TYPHON simulation started from the same crystal structure (blue line). The secondary structure elements are indicated by blue circles for α-helices and red squares for β-strands. The lid region is indicated by the black bracket.

TYPHON used the crystal structure as sole input, from which 581 $C_\alpha \cdots C_\alpha$ Gaussian distance restraints were derived (see Experimental procedures). The sampling ran for 50 million iterations. Figure 1 shows RMS fluctuation calculated from the experimental B-factors for the crystal structure and from 1000 sampled conformations, chosen with a regular interval. The overall flexibility along the sequence is captured well. The lid region clearly displays a higher level of flexibility, in correspondence with its dynamic nature (Skøt et al. 2009). The good agreement with the experimental measure is also reflected in the Pearson correlation coefficient, which is equal to 0.71.

Generating a native ensemble

Advances in nuclear magnetic resonance spectroscopy (NMR) over the past decades made more detailed studies of dynamics in proteins possible. The S^2

order parameter is a measure arising from NMR experiments describing the amplitude of motion of an N-H vector (Lipari and Szabo 1982). A backbone segment that is unrestricted in its movement, usually in a region of high flexibility, will have a low S^2 value. For segments in more constrained or rigid regions of the protein, the S^2 value will be higher. Analyzing S^2 order parameters provides a more direct view on the dynamics of a protein compared to the B-factors. In this test, we analyze whether TYPHON is able to capture the fast dynamics of a protein as implied by the S^2 order parameters.

Ubiquitin is a well studied protein in terms of its dynamics; its relatively small size of 76 amino acids allows for both extensive MD simulations as well as NMR studies. Ubiquitin consists of a five stranded, twisted, antiparallel β-sheet with an α-helix lying across. A number of recent publications discuss the molecular recognition mechanisms using ubiquitin as a model system (Lange et al. 2008; Wlodarski and Zagrovic 2009; Long and Brüschweiler 2011).

TYPHON sampling started from a single crystal structure of ubiquitin (PDB: 1ubi) (Ramage et al. 1994), with 46 automatically detected hydrogen bonds as restraints, and ran for 50 million iterations. A total of 1000 structures were sampled in regular intervals. We also generated an ensemble of 1000 structures using tCONCOORD, starting from the same ubiquitin crystal structure and using default settings. For further comparison, we also included the order parameters calculated from an MD simulation of ubiquitin (Maragakis et al. 2008).

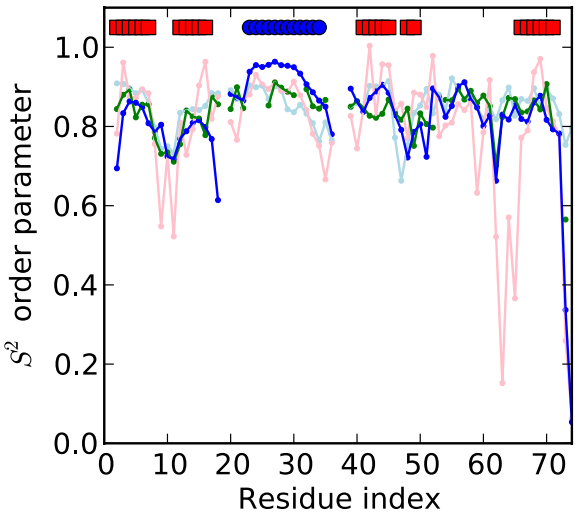


Figure 2: Experimentally determined S^2 values (green) versus values calculated from a TYPHON ensemble (blue) for ubiquitin. The S^2 order parameter is an experimental measure arising from NMR experiments that reflects flexibilities in the protein. It ranges from 0 (isotropic motion) to 1 (no motion). For comparison the figure also shows S^2 order parameters calculated from an MD simulation (light blue) and from a tCONCOORD simulation (light red). The secondary structure is indicated by red squares for β-strands and blue circles for α-helices. The fragmentation of the lines is due to missing values for Ile23, Glu24, Asn25, Gln31, Ile36, Gly53, Arg72, Arg74, Gly75 and Gly76 in the experimental data, and for all proline residues.

Figure 2 shows the S^2 order parameters calculated from the TYPHON ensemble following (Best and Vendruscolo 2004) and order parameters obtained from an experiment (Tjandra et al. 1995). The figure further shows order parameters calculated from a tCONCOORD ensemble obtained with default parameters and from an MD simulation (Maragakis et al. 2008). Overall, the S^2 parameters calculated from the TYPHON ensemble are in good agreement with the experimental measurements; the correlation coefficient for the two curves is 0.73. The most rigid region is located in the well ordered α-helix, between residues 23 and 33. This region is indeed rigid in the TYPHON ensemble as well, though overly so compared to the experimental results (Tjandra et al. 1995). The terminal regions are the most flexible (see Figure 2). Recently, it was found that the increased flexibility in the C-terminus and in loop I, between the β1 and β2 strands, is of importance for the molecular recognition mechanism of ubiquitin (Lange et al. 2008; Wlodarski and Zagrovic 2009). The ensemble generated by TYPHON accurately reflects the conformational fluctuations in these regions of interest.

The order parameters calculated from the MD simulation match the experimental values less well; the correlation coefficient is 0.52. While the MD ensemble accurately reflects the flexibilities in loop I, it does not reproduce the fluctuations in the C-terminus well. The S^2 order parameters calculated from the tCONCOORD ensemble match the general trend of the experimental curve. The correlation coefficient is 0.53, which is also lower than for TYPHON. The generated ensemble appears to overemphasize the flexibility in certain loops, including the functionally important loop I – around residues 7 to 10. In addition, loop V — around residues 63 to 65 — shows considerable discrepancy. Leaving out the flexible C-terminal region following (Lindorff-Larsen et al. 2005) results in correlation coefficients equal to 0.50, 0.55 and 0.28 for the MD, TYPHON and tCONCOORD ensembles, respectively. In conclusion, TYPHON matches the experimentally determined order parameters well, indicating that the fast dynamics — as described by the Lipari-Szabo S^2 parameters — are captured well in the generated ensemble.

Residual dipolar couplings (RDCs) probe the bond vector geometry relative to an external magnetic field. Data acquisition in a nematic phase solvent or in the presence of a paramagnetic center can make measurement of RDCs in the solution state possible (Tjandra et al. 1997; Banci et al. 2004). RDCs are anisotropic quantities and thus average out when molecules undergo isotropic rotational diffusion.

For ubiquitin, Cornilescu *et al.* (Cornilescu et al. 1998) obtained six sets of backbone RDCs in a nematic phase solvent based on phospholipid bicelles. The experimental data was obtained from the Biological Magnetic Resonance Data Bank (BMRB entry: 6457) (Ulrich et al. 2007). We used the same TYPHON and tCONCOORD ensembles as in the previous section. Ensemble

	$N-NH$	$CO-NH$	$C_\alpha-H_\alpha$	$N-CO$	$C_\alpha-CO$	$C_\alpha-C_\beta$
Correlation coefficient average RDC TYPHON	0.91	0.90	0.92	0.94	0.93	0.90
Correlation coefficient average RDC tCONCOORD	0.96	0.91	0.96	0.96	0.96	0.97
Correlation coefficient Crystal structure (1UBI)	0.98	0.96	0.93	0.99	0.99	0.97

Table 1: Statistics for the RDC values obtained from the TYPHON and tCONCOORD ensembles of ubiquitin. Rows one and two: correlation coefficients of the TYPHON and tCONCOORD ensembles with the experimental data, respectively. Third row: correlation coefficient between the crystal structure 1UBI and for all six RDC types.

averages were calculated from these ensembles using the procedure described by Showalter and Bruschweiler (Lindorff-Larsen et al. 2005; Showalter and Bruschweiler 2007).

Figure 3 shows experimentally determined $C_\alpha - CO$ RDCs in comparison with RDCs calculated from a TYPHON ensemble. Supplementary Figure 1 additionally shows correlation plots for all RDCs. In general, there is a good correlation between the values obtained from the TYPHON and the experimental data (see Table 1). The agreement with experiment for the TYPHON ensemble is comparable to the tCONCOORD ensemble and the crystal structure (1UBI). However, Q-factors for the TYPHON ensemble (0.37) are larger than for the tCONCOORD ensemble (0.28) and the crystal structure (0.23), suggesting better qualitative agreement of the tCONCOORD ensemble (Lipsitz and Tjandra 2004).

While reproduction of experimental data such as residual dipolar couplings and order parameters serves as sanity check, it is difficult to make quantitative assessment of the physical time scales sampled (Showalter and Bruschweiler 2007). However, collectively, the results suggest that TYPHON samples broader ensembles in some regions of ubiquitin as compared to tCONCOORD. Regions that appear over-stabilized may be attributed to the employed restraints suggesting that TYPHON ensembles can be improved by input of expert knowledge. In view of the excellent structural quality of the generated decoys (compare section Quality of the sampled structures), these observations support the interpretation of TYPHON as a suitable “null model” of conformational fluctuations in proteins for a given set of restraints; given the nonlocal restraints, the probabilistic models of local structure ensure a thorough exploration of the remaining conformational space.

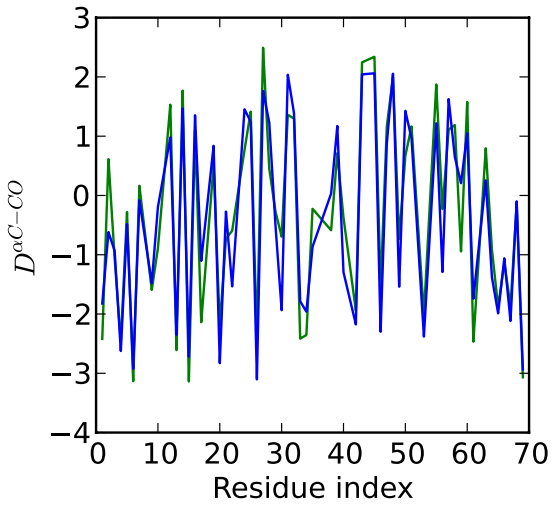


Figure 3: $C_\alpha - CO$ RDC values for ubiquitin. The figure shows a comparison between experimentally determined $C_\alpha - CO$ RDCs (green line) and RDCs calculated from a TYPHON ensemble using the procedure described in (Showalter and Bruschweiler 2007). (blue line), where the RDCs are plotted on the y-axis against the residue index on the x-axis. See also Figure S1.

Functional dynamics of an enzyme

Ribonuclease (RNase) A is a pancreatic protein that cleaves single-stranded RNA; its structural dynamics are essential for its enzymatic function (Doucet et al. 2009; Formoso et al. 2010). The protein has 124 residues and adopts an α/β fold that consists of two domains flanking a catalytic site. In this experiment, we analyze whether TYPHON can reproduce the functional dynamics of RNase A. In addition, we compare the TYPHON ensemble to results obtained from NMA.

We initialized TYPHON sampling from the RNase A crystal structure (PDB: 7RSA) (Wlodawer et al. 1988) and used the automatically detected hydrogen bond network with default settings, resulting in 76 hydrogen bonds and 4 disulfide bridges. The sampling was run for 100 million iterations, from which 1000 structures were retained.

As a measure of the structural flexibility of RNase A, we analyzed 132 experimentally determined structures with a maximum of one point mutation (for a complete list see supplementary Table 1). We superimposed the experimental structures using iterative RMSD minimization to the average structure and calculated the RMSF of the C_α atoms. We call this set the high-sequence similarity PDB ensemble (Best et al. 2006).

In addition, we compare our result to the dynamics of the enzyme according to the elastic network model (ENM), a coarse-grained model of protein dynamics that has been used to analyze collective motions, residue fluctuations, and conformational changes (Tirion 1996; Hinsen 1998; Bahar and Rader 2005; Ma 2005; Kimber et

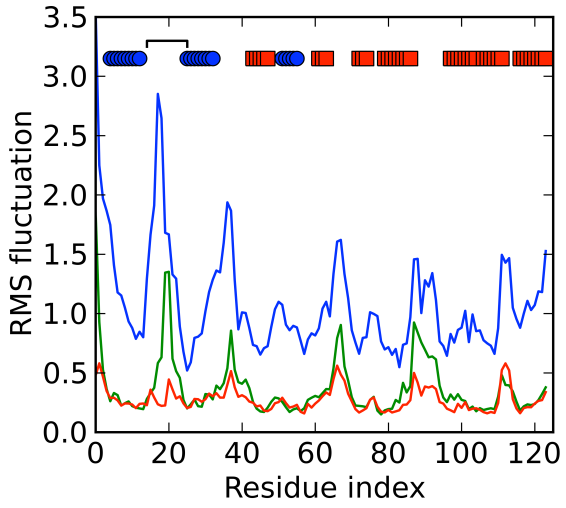


Figure 4: Ribonuclease A dynamics: The plot show the RMS fluctuations measured from a set of PDB structures (see text, green line), from an ENM analysis (red line) and a TYPHON simulation (blue line). The secondary structure elements are indicated by blue circles for helical residues and red squares for strands. Loop I, including residue 14-25, is indicated by the black bracket. See also supplementary Figure 2, supplementary Video SV1 and supplementary Table ST1.

al. 2010). In the ENM, the protein structure is approximated as a network of coupled harmonic oscillators between all C_α atoms closer than a specified cut-off radius. The collective motions of the system can then be calculated using NMA. The ENM analysis was performed with the elNémo server and default parameters, using an 8Å cut-off distance to identify elastic interactions (Suhre and Sanejouand 2004). The server reports the RMSF calculated from the scaled Eigen vectors of the first hundred modes.

The fluctuations found within the PDB and the TYPHON ensembles (Figure 4) are in good agreement; the correlation coefficient is 0.72. The overall flexibility pattern along the amino acid chain indicates increased mobility in the same regions. The amplitude of the fluctuations, however, is significantly larger for the TYPHON ensemble, indicating that a large volume of the conformational space is sampled. This again confirms the interpretation of TYPHON as a suitable null model of conformational fluctuations for a given set of restraints. Notably, loop I - consisting of residues 14 to 25 - has a high degree of flexibility (Figure 4). The dynamics of this loop are especially important for the catalytic activity of the enzyme (Doucet et al. 2009; Formoso et al. 2010). TYPHON sampling started from other crystal structures of RNase A in the PDB yielded similar results (PDB code 3LXO (Doucet et al. 2010) and 2G8Q (Leonidas et al. 2006)). In contrast, while having only a slightly lower correlation coefficient to the PDB ensemble (0.67), the result from the ENM analysis does not show an elevated flexibility in this loop.

The dynamics of loop 1 is a requirement for the functional dynamics of RNase A; RNase has been shown to function through a concerted motion between an open form that can bind substrate and a closed form where catalysis occurs (Watt et al. 2007). In order to investigate how the TYPHON ensemble relates to these motions, we performed a principle component analysis on the TYPHON samples and isolated the main modes. The first mode, which contains the most important variations of the ensemble, indeed shows an opening and closing of the catalytic cleft, lending further evidence that the TYPHON ensemble can be used to explore enzyme dynamics. A video of the motion is part of the supplementary material (see SV1).

Induced change in flexibility

Large scale motions or major changes in flexibility in proteins are often induced by binding or releasing ligands. These ligands can be as complex as multi-atom substrates, inhibitors or drugs or as simple as single metal ions. In this test we use TYPHON to simulate partial unfolding upon loss of metal ions. This application illustrates how the probabilistic models “step in” to provide information in the absence of restraints.

Cu/Zn superoxide dismutase (SOD1) is a ubiquitous protein in the cytoplasm that is associated with the neurodegenerative disease amyotrophic lateral sclerosis (ALS). ALS results in paralysis and respiratory failure within one to five years from onset (Pasinelli and Brown 2006). The oligomerization of SOD1 is associated with a gain in toxic function. Experimental evidence suggests that a loss of the two metal ions induces structural changes to the monomeric form of SOD1 and subsequently leads to pathogenic aggregation (Teilum et al. 2009b). The exact pathway is however still unknown. We used the PDB:2v0a crystal structure as starting point for our experiments (Strange et al. 2007). SOD1 consists of a β barrel with long loops connecting the antiparallel strands. It contains a disulfide bridge and has two associated metal ions: a copper ion that is coordinated by four histidines (residues 46, 48, 63, 120) and a zinc ion that is coordinated by three histidines and an aspartate (residues 63,71, 80 and 82).

Ding and coworkers performed a molecular dynamics analysis of the SOD1 monomer (Ding and Dokholyan 2008). They systematically tested the effect of losing metal ions and/or reducing the disulfide bridge. Each individual event leads to a significant increase in flexibility; the two most affected regions are both located in the long loop IV (Figure 5, A). The region around Cys57, which is involved in the disulfide bridge, is primarily affected by the loss of the disulfide bridge. The loss of the metal ions primarily affects the regions adjacent to the ion coordinating histidines. Other parts of the structure seem mostly unaffected by either event. Following this study, we analyzed the mobility of different forms of SOD1, namely the holo form with the C57-C146 disulfide bridge intact and the apo monomer with the disulfide bridge reduced. Again we used only a single crystal structure as starting point. We set up two different TYPHON experiments. For the apo experiment, we removed the automatically detected disulfide bridge and did not include any restraints involving the metal ions. For the holo experiment, we added the copper and zinc ions in the form of distance restraints that maintain the mutual

distances between the four ion-coordinating atoms (see supplementary Table 2), and included the disulfide bridge. The remaining restraint network, consisting of 73 automatically detected hydrogen bonds, was identical in both setups. For each setup we ran three experiments of 100 million iterations each, and combined the generated structures for the final evaluation in order to ensure converged sampling. Note that in the absence of the restraints concerning metal ions and disulfide bridge, the relative influence of the probabilistic models of local structure on the sampled conformations increases.

Figure 5 shows the results of the different experiments in putty representation. The results show that the loss of the metal ions and the reduced disulfide bridge leads to a significant increase in flexibility, especially in the long loop IV between residues 49 and 83, but also in the loops II, VI and VII. The spike in flexibility around residue 57 can be attributed to the reduced disulfide bridge, which in the native structure covalently binds this surface loop. The increased flexibility in other parts of the protein is likely due to the loss of the metal ions. An interesting observation is also the increased flexibility in loop II around residue 25, which is not in direct contact with any of the mutated sites. We speculate that the overall increased mobility in the long loop IV and VI also influenced the flexibility in this region.
The results closely resemble the results of Ding *et al.* that were obtained from molecular dynamics simulations. A TYPHON experiment requires about twenty hours, which would allow scanning of larger sets of clinically known mutations (Andersen et al. 2003). We point out that the increased mobility in loop II was not observed in the MD study of Ding *et al. (Ding and Dokholyan 2008)*, which illustrates that TYPHON can deliver results that suggest starting points for new hypotheses or follow up studies. It should be noted that TYPHON only includes the steric component of the ion loss; changes in electrostatics or solvent accessibility are not directly accounted for. Nonetheless, in this case, modeling the effect of the metal ions as simple Gaussian restraints accurately reproduces the results obtained from much more sophisticated simulations, and leads to potentially interesting new observations.

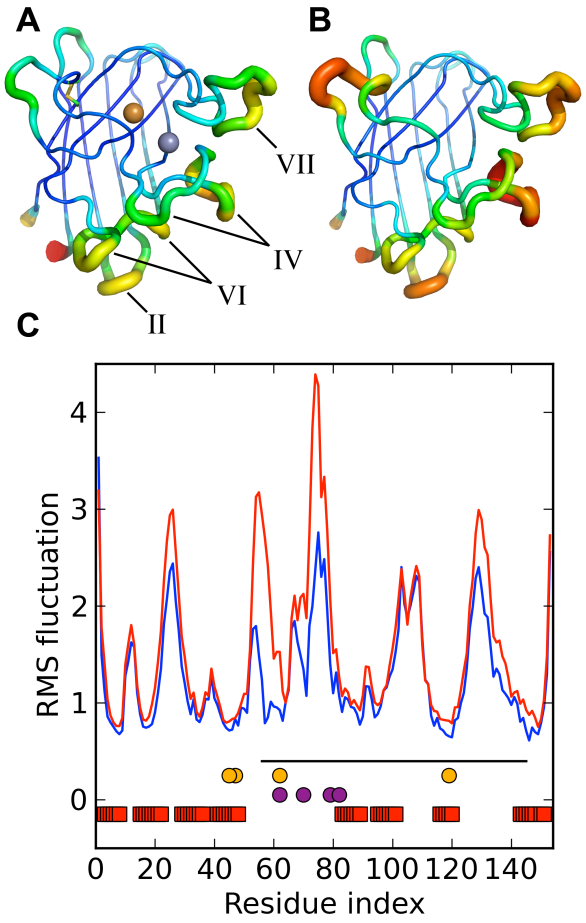


Figure 5: Cu/Zn Superoxide Dismutase (SOD1). (A) TYPHON ensemble obtained from the native monomer. The Cu and Zn ions are shown as an orange and a purple sphere, respectively. The C57-C146 disulfide bridge is shown as a stick representation. The roman numerals indicate the loop numbers. This ensemble corresponds to the blue line in (C). (B) TYPHON ensemble obtained without the ions and with the disulfide bridge reduced. This ensemble corresponds to the red line in (C). Panel (C) shows the corresponding RMSF curves. The disulfide bridge is indicated as a black line. The residues coordinating the metal ions are marked by orange and purple circles for the copper and zinc ion, respectively.

Local structure under the control of probabilistic models

The Gaussian restraints obviously do not allow for formation or dissolution of nonlocal interactions; the restraint network is rigorously fixed during the sampling procedure. However, certain nonlocal interactions, such as hydrogen bonds in helices, can be put under the control of the probabilistic models instead. In practice, this means that certain conformational fluctuations of the protein backbone on a local length scale could be investigated. In this application, we explore and illustrate this approach with a small helical protein and investigate helical mobility and $\alpha / 3_{10}$ -helix transitions.

The Mature-T-Cell Proliferation Gene 1 (MTCP1) is a known oncogene that is linked to certain types of leukemia (Barthe et al. 2002). The structure of the human $p8^{MTCP1}$ protein has been solved by NMR and consists of three helices. A stable α hairpin connecting helix I and II is covalently held together by two disulfide bridges between residues 7, 38 and 17, 28 respectively. A third, less restricted and stable helix (helix III) is also connected to helix II with a third disulfide bridge between residues 39 and 50 (Barthe et al. 1997). MD simulations indicate that helix III is fairly flexible with respect to the α hairpin (Barthe et al. 2002).

We first investigate to which extent the helices move with respect to each other. We therefore started from the first model of a $p8^{MTCP1}$ NMR ensemble (PDB: 2hp8) (Barthe et al. 1997). The experiment ran for 100 million iterations with the three disulfide bridges as only restraints. However, we also imposed the secondary

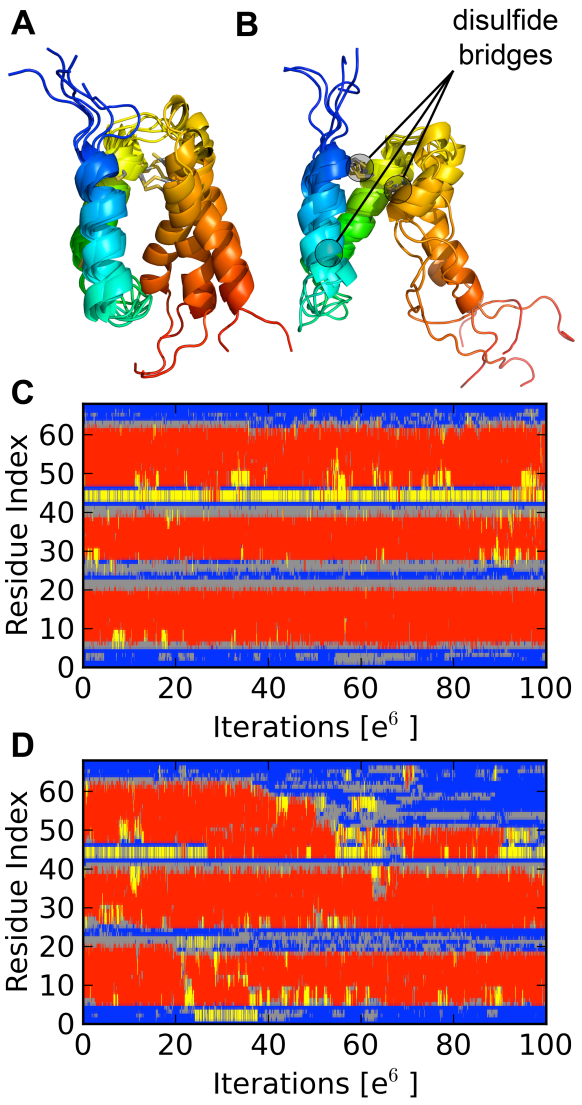


Figure 6: Local structure under the control of probabilistic models. (A,B) Five representative structures of the simulation (A) with and (B) without fixed secondary structure assignment. The disulfide bridges are shown in stick representation and highlighted in panel (B). (C,D) Secondary structure content of the simulation (C) with and (D) without fixed secondary structure input. The secondary structure was measured using DSSP. Color code: red is α-helix; yellow is 3_{10}-helix; gray is β-turn; blue is random coil.

structure of the native structure according to DSSP (Kabsch and Sander 1983) through TorusDBN (Boomsma et al. 2008). This is a more flexible and "soft" way to restrain the sampling, as the helical regions are allowed to bend, or to a certain extent form and dissolve hydrogen bonds under the influence of the probabilistic model.

Despite the absence of restraints besides those involving the three disulfide bridges, all helices remain stable throughout the sampling. Figure 6 A shows five representative structures from the ensemble. Helix I and helix II are tightly fixed by the interhelical disulfide bridges, which only allow limited movements. Helix III is only tethered by a single disulfide bond in the beginning of the helix, which results in higher flexibility. As indicated in Figure 6 A, helix III slightly tilts away from the other two helices, a behavior that has also been observed in MD simulations (Barthe et al. 2002).

Figure 6 C shows the secondary structure content over the course of the first experiment. The consistent red bars show that all three helices remain fully helical throughout the sampling. In the beginning of helix III, we observe transitions between α- and 3_{10}-helix, which is again in agreement with the results of an MD simulations (Barthe et al. 2002).

In the second experiment, we investigate the stability of the helices themselves. We again included restraints concerning the three native disulfide bonds. However, this time we did not provide any secondary structure information to TorusDBN. In other words, this means that TorusDBN still enforces protein-like conformations, but does not require them to be helical.

Again helix I and helix II remain stable throughout the sampling as indicated by the consistent red bars in Figure 6 D. This is not surprising since both helices are covalently connected near their respective start and end. The entire protein structure however, is significantly more flexible, expressed by the movement of the helices with respect to each other (compare Figure 6 B). In contrast to helices I and II, helix III quickly unfolds up to residue 50, where it is covalently attached to helix II via a disulfide bridge.

In addition to the unfolding helix III, we observe significant differences compared to the first experiment in the loop regions. In particular, for loop II, which connects helix II and III and stretches from residue 39 to 47, we observe a transition to an α-helix. The terminal 18 residues of helix III readily unfold (see Figure 6 D), which points to a difference in stability between the first two and the third helix.

This experiment strikingly demonstrates the possibilities of probabilistic models. In the first experiment, which includes the disulfide bridges and secondary structure information, we observed specific movements of the helices with respect to each other, and transitions from an α- to a 3_{10}-helix in the beginning of helix III. Both observations concur with the results obtained from MD simulations (Barthe et al. 2002). In the second experiment, which includes the disulfide bridges but not the secondary structure information, we obtained some information on the relative stability of the helices themselves. Helices I and II remain stable, while helix III readily unfolds. Again, this difference in stability is in accordance with MD simulations (Barthe et al. 2002).

Quality of the sampled structures

In order to evaluate the quality of the structures, we analyzed 50 random structures from an RNase A ensemble, generated as described above, using PROCHECK (Laskowski et al. 1993). For comparison, we generated 50 tCONCOORD (Seeliger et al. 2007) samples for the same protein (starting from PDB: 7rsa).

Residues in regions	tCONCOORD	TYPHON
Most favored	69.8%	88.1%
Additionally allowed	26.3%	10.8%
Generously allowed	3.1%	0.7%
Disallowed	0.7%	0.5%
ϕ/ψ G factor	-0.93	-0.24
χ_1 G factor	-0.26	0.10
Overall G factor	-0.69	-0.13

Table 2: Quality assessment of structures generated by TYPHON. The table lists the results of a PROCHECK analysis of a set of TYPHON and tCONCOORD samples. Well-refined structures usually have 90% or more percent of all residues in the most favored regions. The G factor is a log odds score; higher numerical values denote higher quality. See also supplementary documents SD 1 and SD2.

The detailed PROCHECK reports can be accessed as supplementary documents SD1 and SD2.

The Ramachandran map divides the main chain's conformational space, as parameterized by the ϕ and ψ angles, in different regions, some sterically more favorable then others (Ramachandran et al. 1963). Well-refined protein structures are expected to have 90% or more of the backbone dihedral angles in the most favorable regions. The PROCHECK analysis indicates that the TYPHON samples are of good quality; over 88% of all angles are in the most favored regions. In contrast, the tCONCOORD samples have less than 70% of the backbone angles in these favored regions (Table 2).

PROCHECK's G factor is a measure of how well the analyzed structures match the observed distributions of bond lengths, bond angles and dihedral angles in crystal structures and is expected to be -0.5 or higher for well-refined structures. Also in this respect, TYPHON samples have a higher quality then tCONCOORD samples; the G-factor is -0.13 versus -0.69. The G factor takes the side chain quality into account; in this respect, TYPHON undoubtedly benefits from the detailed side chain modeling in BASILISK (Harder et al. 2010).

Additionally, we performed a WHATIF (Vriend 1990) packing analysis of the TYPHON and tCONCOORD ensembles of RNase A. The structures generated by TYPHON have an average packing environment score of -1.495. Those generated by tCONCOORD have an average score of -1.944. As well-refined structures have a score around -0.5, both methods might be improved in this respect.

Computational efficiency

TYPHON is computationally efficient. The ubiquitin experiments used in this study were performed on a regular desktop computer (Intel Core i7, 2.8GHz) and ran for around 10 hours on a single CPU core. The human $p8^{MTCP1}$ protein experiments comprising 100 million iterations ran for about 15h. Naturally, the runtime increases as the number of restraints in the network grows, though extensive caching in the calculations minimizes this effect to an extent. With increasing protein size, more iterations will be necessary to achieve a comparable level of convergence. While a parallelization of a single run onto multiple cores is not possible in the current implementation, it is possible to perform several TYPHON experiments in parallel to obtain better statistics.

Discussion

In this paper we present TYPHON, a novel approach to explore conformational fluctuations in proteins. TYPHON incorporates detailed probabilistic models of the conformational space of a protein's main chain and its amino acid side chains (Boomsma et al. 2008; Harder et al. 2010) and an efficient local backbone resampling algorithm (Bottaro et al. 2012). During sampling by TYPHON, the conformational space is restricted by a set of restraints imposed on the structure. These restraints typically concern nonlocal interactions such as hydrogen bonds, disulfide bridges or interactions with metal ions. The protein structure on a local length scale, including main chain and side chains, is controlled by the probabilistic models.

In this study, we show that TYPHON is able to generate structural ensembles that closely resemble native ensembles described by experimental measures. This includes fluctuations as measured by S^2 order parameters, as well as measured by RDC values. The RNase A study shows not only that TYPHON captures the functional dynamics in the correct regions, but also that a principal component analysis of the results is feasible to identify large-scale motions. The analysis of the superoxide dismutase results shows that TYPHON can be used to model effects due to the gain or loss of a ligand, including partial unfolding.

Its computational efficiency makes TYPHON a promising tool for larger screening efforts; for example, of known mutations with clinical relevance. Another interesting application lies in generating suitable candidate structure for docking experiments, allowing for some degree of flexibility in the binding pocket (Henzler and Rarey 2010). The high quality of the generated structures indicates that no irrelevant parts of the conformational space are explored. On the other hand, TYPHON thoroughly samples the relevant conformational space.

The results of the human $p8^{MTCP1}$ protein experiments demonstrate another strength of our approach. With only a minimal set of restraints defined for the system, the effect of the probabilistic models becomes obvious. They control the local structure and maintain the overall secondary structure, while still allowing for significant conformational fluctuations. It should be noted that it is also possible to run TYPHON without explicitly defining the secondary structure, leading to significantly broader sampling.

In the current implementation, TYPHON keeps the constraint network fixed during the sampling. As a next step, it would be advantageous to allow more flexibility in the restraint network, such as the dissolution or formation of arbitrary hydrogen bonds as the sampling progresses. However, this will require the development of a suitable probabilistic model of nonlocal interactions in proteins, and its seamless combination with the probabilistic models of local structure. Fortunately, important theoretical progress was recently made in this respect (Hamelryck et al. 2010). Another interesting addition would be to directly include restraints from experimental data (Olsson et al. 2011).

Availability

TYPHON is available as part of the Phaistos package (Borg et al. 2009) and can be obtained freely from sourceforge under the GNU public license[1]. Currently the Phaistos package is limited to single chain proteins. However, support for multiple chains will be added in the next release.

Acknowledgments

The authors thank Francesco Carbone for help with the $p8^{MTCP1}$ study. We thank Kaare Teilum (University of Copenhagen), Kresten Lindorff-Larsen (University of Copenhagen), Thomas Poulsen (Novozymes) and Leonardo De Maria (Novozymes) for valuable comments and suggestions. T. Harder and M. Borg are funded by the Danish Council for Strategic Research (NABIIT, 2106-06-0009). W. Boomsma and S. Olsson are funded by the Danish Council for Independent Research (FNU, 272-08-0315 and FTP, 274-09-0184, respectively). S. Bottaro acknowledges funding from Radiometer, DTU Elektro.

EXPERIMENTAL PROCEDURES

Overview

The TYPHON network calculation starts from a full atom protein structure, including all the hydrogen atoms. A restraint network is either loaded from an input file or created according to the protocol described in the following section. In the course of the sampling, the dihedral angles in both the main chain and as well as in the side chains are modified under the control of TorusDBN (Boomsma et al. 2008) and BASILISK (Harder et al. 2010), respectively. An efficient local moves method makes subtle movements of the protein backbone possible (Bottaro et al. 2012), and also affects the bond angles in the backbone (see below).

Restraint network calculation

TYPHON currently supports three classes of restraints, involving hydrogen bonds, disulfide bridges and distance restraints between arbitrary atoms. In the absence of any user input, the program suggests a network using default parameters described in the following paragraphs. This default network is mainly based on biologically relevant restraints, such as hydrogen bonds and disulfide bridges. In order to stabilize parts of the protein that are naturally stabilized by effects that are not modeled explicitly, TYPHON also connects residues that are far apart in the amino acid sequence, but close in space. The user can edit the generated network by adding, removing or modifying restraints between arbitrary atoms.

We evaluate all potential hydrogen bonds using the DSSP hydrogen bond energy (Kabsch and Sander 1983). Following Kabsch and co-workers, we discard all candidates with a DSSP energy higher than $-0.5\frac{\text{kcal}}{\text{mol}}$. If an atom has multiple potential hydrogen bonding partners, only the one with the lowest energy is retained. Following the general idea of the DSSP hydrogen bond energy, the hydrogen bond geometry is modeled using four distances. For backbone-backbone hydrogen bonds, these respective distances are explained in more detail in supplementary Figure 3. For hydrogen bonds involving side chains, the corresponding standard hydrogen bond acceptors and donors are used; asparagine, aspartate, glutamine and glutamate can act as hydrogen bond acceptors; arginine, asparagine, glutamine, histidine, lysine, serine, threonine, tryptophan and tyrosine can act as hydrogen bond donors.

Disulfide bridges are required to have a $S_\gamma \cdots S_\gamma$ distance of 3 Å or less. Similar to hydrogen bonds, the geometry of the disulfide bond is also modeled by four distances, consisting of the $S_\gamma \cdots S_\gamma$, $C_\beta \cdots S_\gamma$, $S_\gamma \cdots C_\beta$ and $C_\beta \cdots C_\beta$ distances.

The last class of restraints that are detected by default connects residues that are far apart in the amino acid sequence but close together in space. These restraints stabilize parts of the protein that are naturally stabilized by effects not accounted for explicitly in TYPHON, such as hydrophobic interactions. Residue pairs that are five or more residues apart in the sequence but within six Å ($C_\alpha \cdots C_\alpha$ distance) are modeled by a Gaussian probability distribution on the distance between the two C_α atoms. The distance in the input structure is used as mean μ. The variance σ^2 is set proportional to the square of the distance:

$$\sigma^2 = \left(\frac{\mu}{6}\right)^2$$

This value was chosen by trial-and-error and produces reasonable results. It allows for more flexibility with increasing distance.

[1] http://sourceforge.net/projects/phaistos/

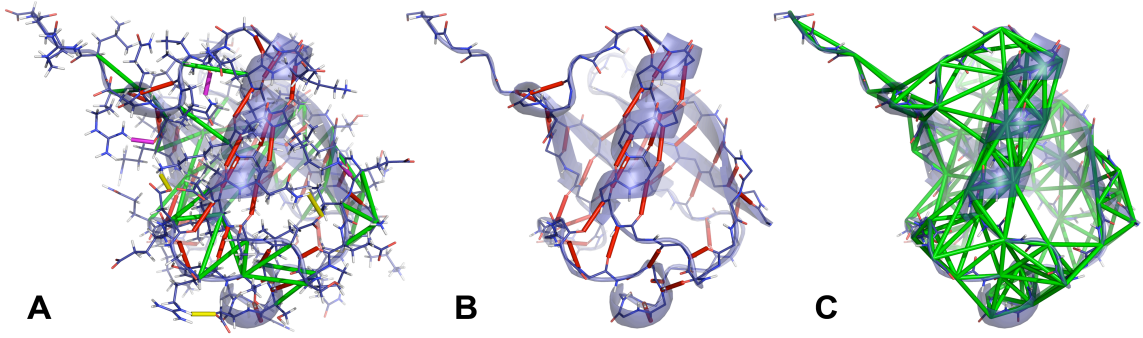


Figure 7: Restraint Network. Depicted are three different calculated networks for ubiquitin (PDB 1ubi). (A) A network that includes all hydrogen bond types (red: backbone hydrogen bonds, purple: backbone-side chain hydrogen bonds, yellow: side chain-side chain hydrogen bonds) as well as C_α contacts (green). (B) A network that includes only the backbone hydrogen bonds. (C) A network that only includes C_α contacts. The cutoff was 7 Å. The minimum sequence separation between the residues in the chain was two. See also supplementary figure 3.

Especially when modeling large scale movements, the automatically detected restraints will not always yield the best results. In order to keep the framework flexible and to utilize the expert knowledge of the researcher, TYPHON allows modifying the restraints and adding distance restraints between arbitrary atom pairs in the structure. In that way, the researcher may additionally stabilize certain parts of the structure or allow more flexibility in other parts. It is also possible to remove automatically detected restraints, for example when a certain hydrogen bond is known to be weak.

To further simplify this process, TYPHON can generate a PyMOL (Schrödinger 2010) script that visualizes the restraints. This makes it possible to quickly detect regions that need manual, expert interaction. Figure 7 shows different restraint networks visualized using the generated PyMOL script.

Unstable hydrogen bonds

Hydrogen bonds that are in direct contact with solvent molecules are known to be significantly less stable than those that are well shielded. Fernandez *et al*. (Fernández and Berry 2002; Fernández et al. 2002; Fernández and Scott 2003; Fernández 2010) proposed the concept of *dehydrons*; insufficiently shielded hydrogen bonds that are more likely to break. They showed that the number of the carbonaceous groups, CH_n, in a shell around the hydrogen bond is a good estimate of water accessibility. tCONCOORD incorporates this convenient measure to judge the stability of a hydrogen bond (Seeliger et al. 2007). We extended their approach, which was only formulated for backbone hydrogen bonds, to apply to hydrogen bonds involving side chains as well. We therefore moved the centers of the two spheres composing the dehydration shell to the donor nitrogen and the acceptor carbon atoms (Fernández and Berry 2002; Fernández et al. 2002; Fernández and Scott 2003; Fernández 2010). We recalibrated the measure using counts of carbonaceous groups derived from a set of high resolution crystal structures previously used as training data for BASILISK (Harder et al. 2010). Following Fernandez *et al.* (Fernández and Berry 2002; Fernández et al. 2002; Fernández and Scott 2003; Fernández 2010), we defined the threshold between weak and strong hydrogen bonds as the 4% percentile of the counts. This resulted in thresholds equal to 14, 9 and 7 for backbone-backbone, backbone-side chain and side chain-side chain hydrogen bonds, respectively. All weak hydrogen bonds are removed from the restraint network by default.

Protein backbone move

TYPHON sampling is usually started from the native state of a protein, that is, from a densely packed, compact structure. In order to capture the subtle movements and flexibilities in compact proteins, it is important to propose local updates of the backbone conformation. A *local move* only affects a limited part of the protein backbone - such as a stretch of five residues - while the rest of the protein remains unchanged.

In TYPHON, we use a novel type of local move, called CRISP (Bottaro et al. 2012). Similar to other methods (Go and Scheraga 1970; Dodd et al. 1993; Hoffmann and Knapp 1996; Ulmschneider and Jorgensen 2003), a local move consists of a concerted rotation of the bond and dihedral angles of the backbone atoms of neighboring residues. Each move involves four elementary steps: (1) Choose a random stretch in the protein chain. (2) *Pre-rotation*: Propose a set of bond and dihedral angle variations in the first $N-6$ degrees of freedom. (3) *Post-rotation*: calculate the six remaining degrees of freedom such that the loop closes. (4) Calculate the bias introduced by performing such a non-random modification of the chain. The bias calculation is important when the method is used in a Markov chain Monte Carlo sampling scheme to ensure detailed balance.

This geometrical problem is tackled in a novel and original manner. We derived an analytical solution for the post-rotational step, thus avoiding the tedious numerical solution of a system of six equations for the six unknown degrees of freedom. The analytical solution is used to

derive an efficient strategy to draw tentative updates of the chain. This scheme makes it possible to continuously control the angular variations of all degrees of freedom involved. The CRISP method thus improves on previous concerted-rotation methods where, in order to satisfy all geometrical restraints, tentative updates of the chain are often radically different from the original structure or introduce a suboptimal local structure.

Protein side chain move

To propose a new side chain conformation, we use our previously developed probabilistic model of side chain conformational space, BASILISK (Harder et al. 2010). BASILISK is a dynamic Bayesian network that makes it possible to sample side chain conformations for all relevant amino acids in continuous space. By default, TYPHON resamples a single, randomly picked residue at a time, proposing an entirely new set of χ angles for the side chain. Both the bond length and the bond angles remain unchanged. In order to have a roughly equal amount of accepted changes affecting side chains and backbone, TYPHON on average resamples five side chains for every backbone move, since a local move affects five backbone residues.

Sampling strategy and scoring functions

For sampling, we use a classic Markov chain Monte Carlo (MCMC) approach. According to the Metropolis-Hastings (Metropolis et al. 1953; Bishop 2006) sampling scheme, a newly proposed X' structure is accepted with the following likelihood:

$$P_{\text{acc}}(X \to X') = \min\left(1, \frac{P(X')Q(X' \to X)}{P(X)Q(X \to X')}\right)$$

where $P_{\text{acc}}(X \to X')$ is the probability of accepting to move from structure X to structure X'; $P(X)$ and $P(X')$ is the probability of X and X', respectively; $Q(X \to X')$ and $Q(X' \to X)$ are the probabilities of proposing to move from X to X' and from X' and X, respectively. $P(X)$ is defined as:

$$P(X) \propto P_{\text{R}}(X)P_{\text{T}}(X \mid A)P_{\text{B}}(X \mid A)\Delta(X)$$

where A is the amino acid sequence; $P_{\text{R}}(X)$ is the probability density of the restraint network, consisting of the product of the probability densities of the individual Gaussian restraints; $P_{\text{T}}(X \mid A)$ is the density of the backbone angles according to TorusDBN; $P_{\text{B}}(X \mid A)$ is the probability density of the side chain angles according to BASILISK; and $\Delta(X)$ is a clash term that is either one or zero. This simple clash function is introduced to avoid close contacts between atoms. We reject every structure with one or more atom pairs below a specific distance cutoff. The exact cutoff distance depends on the atoms involved: 1.5 Å for a hydrogen atom and any other atom; 1.8 Å for S_γ atoms, in order to allow disulfide bridges; and 2.3 Å for any other atom pair.

The proposal distributions consist of resampling of side chain conformations using BASILISK (Harder et al. 2010), or local moves using CRISP. To facilitate smooth local perturbations of the backbone chain, CRISP allows for small variations of the backbone bond-angles. Each angle is modeled by an atom specific Gaussian distribution with parameters chosen in accordance with the bond-angle term of the OPLS-AA force field (Jorgensen et al. 1996; Kaminski et al. 2001).

REFERENCES

Andersen, P. M., K. B. Sims, et al. (2003). "Sixteen novel mutations in the Cu/Zn superoxide dismutase gene in amyotrophic lateral sclerosis: a decade of discoveries, defects and disputes." Amyotroph Lateral Scler **4**(2): 62-73.

Bahar, I. and A. J. Rader (2005). "Coarse-grained normal mode analysis in structural biology." Curr Opin Struct Biol **15**: 586-592.

Banci, L., I. Bertini, et al. (2004). "Paramagnetism-based restraints for Xplor-NIH." J Biomol NMR **28**: 249-261.

Barthe, P., C. Roumestand, et al. (2002). "Helix motion in protein C12A-p8^{MTCP1} Comparision of molecular dynamics simulations and multifield NMR relaxation data." J Comp Chem **23**: 1577-1587.

Barthe, P., Y. S. Yang, et al. (1997). "Solution structure of human p8^{MTCP1}, a cysteine-rich protein encoded by the MTCP1 oncogene, reveals a new alpha-helical assembly motif." J Mol Biol **274**: 801-815.

Berman, H. M., J. Westbrook, et al. (2000). "The protein data bank." Nucleic Acids Res **28**: 235-242.

Best, R. and M. Vendruscolo (2004). "Determination of protein structures consistent with NMR order parameters." J Am Chem Soc **126**(26): 8090-8091.

Best, R. B., K. Lindorff-Larsen, et al. (2006). "Relation between native ensembles and experimental structures of proteins." Proc Natl Acad Sci U S A **103**(29): 10901-10906.

Bishop, C. M. (2006). Pattern recognition and machine learning, Springer.

Boomsma, W., K. V. Mardia, et al. (2008). "A generative, probabilistic model of local protein structure." Proc Natl Acad Sci U S A **105**: 8932-8937.

Borg, M., K. V. Mardia, et al. (2009). "A probabilistic approach to protein structure prediction: PHAISTOS in CASP9." LASR 2009: 65-70.

Bottaro, S., W. Boomsma, et al. (2012). "Subtle Monte Carlo updates in dense molecular systems." J Chem Theory Comput **8**(2): 695-702.

Brooks, B. R., C. L. Brooks III, et al. (2009). "CHARMM: The biomolecular simulation program." J Comp Chem **30**: 1545-1615.

Cornilescu, G., J. L. Marquardt, et al. (1998). "Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase." J Am Chem Soc **120**: 6836-6837.

de Groot, B. L., D. M. van Aalten, et al. (1997). "Prediction of protein conformational freedom from distance constraints." Proteins **29**: 240-251.

Ding, F. and N. V. Dokholyan (2008). "Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation." Proc Natl Acad Sci U S A **105**(50): 19696-19701.

Dobson, C. M. (2003). "Protein folding and misfolding." Nature **426**: 884-890.

Dodd, L., T. Boone, et al. (1993). "A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses." Mol Phys **78**: 961-996.

Doucet, N., T. B. Jayasundera, et al. (2010). "The crystal structure of ribonuclease A in complex with thymidine-3 '-monophosphate provides further insight into ligand binding." Proteins **78**(11): 2459-2468.

Doucet, N., E. D. Watt, et al. (2009). "The flexibility of a distant loop modulates active site motion and product release in Ribonuclease A." Biochemistry **48**: 7160-7168.

Fernández, A. (2010). Transformative concepts for drug design: Target wrapping, Springer, Heidelberg.
Fernández, A. and S. Berry (2002). "Extend of hydrogen-bond protection in folded protein: a constraint on packing architectures." Biophys J **83**: 2475-2481.
Fernández, A. and R. Scott (2003). "Dehydron: a structurally encoded signal for protein interaction." Biophys J **85**(3): 1914-1928.
Fernández, A., T. Sosnik, et al. (2002). "Dynamics of hydrogen bond desolvation in protein folding." J Mol Biol **321**: 659-675.
Formoso, E., J. M. Matxain, et al. (2010). "Molecular dynamics simulation of bovine pancreatic Ribonuclease A-CpA and transition state-like complexes." J Phys Chem B **114**(21): 7371-7382.
Go, N. and H. Scheraga (1970). "Ring closure and local conformational deformations of chain molecules." Macromolecules **3**: 178-187.
Hamelryck, T., M. Borg, et al. (2010). "Potentials of mean force for protein structure prediction vindicated, formalized and generalized." PloS One **5**(11): e13714.
Hamelryck, T., J. T. Kent, et al. (2006). "Sampling realistic protein conformations using local structural bias." PLoS Comput Biol **2**: e131.
Harder, T., W. Boomsma, et al. (2010). "Beyond rotamers: a generative, probabilistic model of side chains in proteins." BMC bioinformatics **11**: 306.
Henzler, A. M. and M. Rarey (2010). "In pursuit of fully flexible protein-ligand docking: Modeling the bilateral mechanism of binding." Mol Inform **29**(3): 164-173.
Hess, B., C. Kutzner, et al. (2008). "Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation." J Chem Theory Comput **4**(3): 435-447.
Hinsen, K. (1998). "Analysis of domain motions by approximate normal mode calculations." Proteins **33**(3): 417-429.
Hoffmann, D. and E. W. Knapp (1996). "Protein dynamics with off-lattice Monte Carlo moves." Phys Rev E **53**: 4221-4224.
Jacobs, D. J., A. J. Rader, et al. (2001). "Protein flexibility predictions using graph theory." Proteins **44**(2): 150-165.
Jorgensen, W. L., D. S. Maxwell, et al. (1996). "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids." J Am Chem Soc **118**: 11225-11236.
Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**: 2577-2637.
Kaminski, G. A., R. A. Friesner, et al. (2001). "Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides." J Phys Chem B **105**: 6474-6487.
Karplus, M. and J. A. McCammon (2002). "Molecular dynamics simulations of biomolecules." Nat Struct Biol **9**(9): 646-652.
Kimber, M. S., A. Y. H. Yu, et al. (2010). "Structural and theoretical studies indicate that the cylindrical protease ClpP samples extended and compact conformations." Structure **18**(7): 798-808.
Klepeis, J. L., K. Lindorff-Larsen, et al. (2009). "Long-timescale molecular dynamics simulations of protein structure and function." Curr Opin Struct Biol **19**(2): 120-127.
Kuzmanic, A. and B. Zagrovic (2010). "Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors." Biophys. J. **98**(5): 861.
Lange, O. F., N. A. Lakomek, et al. (2008). "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution." Science **320**: 1471-1475.
Laskowski, R. A., M. W. MacArthur, et al. (1993). "PROCHECK: a program to check the stereochemical quality of protein structures." J Appl Cryst **26**: 283-291.
Leonidas, D. D., T. K. Maiti, et al. (2006). "The binding of 3 '-N-piperidine-4-carboxyl-3 '-deoxy-ara-uridine to ribonuclease A in the crystal." Bioorg Med Chem **14**(17): 6055-6064.
Levitt, M., C. Sander, et al. (1983). "The normal modes of a protein: Native bovine pancreatic trypsin inhibitor." Int J Quantum Chem **24**: 181-199.
Lindorff-Larsen, K., R. B. Best, et al. (2005). "Simultaneous determination of protein structure and dynamics." Nature **433**(7022): 128-132.
Lipari, G. and A. Szabo (1982). "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results." J Am Chem Soc **104**(17): 4559-4570.
Lipsitz, R. S. and N. Tjandra (2004). "Residual dipolar couplings in NMR structure analysis." Annu Rev Biophys Biomol Struct **33**: 387-413.
Long, D. and R. Brüschweiler (2011). "In silico elucidation of the recognition dynamics of ubiquitin." PLoS Comput Biol **7**(4): e1002035.
Ma, J. (2005). "Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes " Structure **13**(3): 373-380.
Maragakis, P., K. Lindorff-Larsen, et al. (2008). "Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins." J Phys Chem B **112**(19): 6155-6158.
Mardia, K. V. and P. E. Jupp (2000). Directional statistics, John Wiley and Sons, New York, USA.
Marrink, S. J., H. J. Risselada, et al. (2007). "The MARTINI force field: coarse grained model for biomolecular simulations." J Phys Chem B **111**: 7812-7824.
McCammon, J. A., B. R. Gelin, et al. (1977). "Dynamics of folded proteins." Nature **267**: 585-590.
Metropolis, N., A. W. Rosenbluth, et al. (1953). "Equations of State Calculations by Fast Computing Machines." J. Chem. Phys. **21**(6): 1087–1092.
Olsson, S., W. Boomsma, et al. (2011). "Generative probabilistic models extend the scope of inferential structure determination." J Magn Reson **213**(1): 182-186.
Pasinelli, P. and R. H. Brown (2006). "Molecular biology of amyotrophic lateral sclerosis: insights from genetics." Nat Rev Neurosci **7**: 710-723.
Ramachandran, G. N., C. Ramakrishnan, et al. (1963). "Stereochemistry of polypeptide chain configurations." J Mol Biol **7**: 95-99.
Ramage, R., J. Green, et al. (1994). "Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin." Biochem J **299**: 151-158.
Schrödinger, L. (2010). The PyMOL Molecular Graphics System, Versio1 1.3r1.
Seeliger, D. and B. L. De Groot (2010). "Conformational transitions upon ligand binding: Holo-structure prediction from apo conformations." PLoS Comput. Biol **6**(1): e1000634.
Seeliger, D. and B. L. D. Groot (2009). "tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules." J Comput Chem **30**(7): 1160-1166.
Seeliger, D., J. Haas, et al. (2007). "Geometry-based sampling of conformational transitions in proteins." Structure **15**(11): 1482-1492.
Shaw, D. E., P. Maragakis, et al. (2010). "Atomic-level characterization of the structural dynamics of proteins." Science **330**(6002): 341-346.
Showalter, S. A. and R. Bruschweiler (2007). "Quantitative molecular ensemble interpretation of NMR dipolar couplings without restraints." J Am Chem Soc **129**(14): 4158-+.
Skøt, M., L. de Maria, et al. (2009). "Understanding the Plasticity of the α/β Hydrolase Fold: Lid Swapping on the Candida antartica Lipase B Results in Chimeras with Interesting Biocatalytical Properties." BioChemBio **10**: 520-527.
Strange, R. W., C. W. Yong, et al. (2007). "Molecular dynamics using atomic-resolution structure reveal structural fluctuations that may lead to polymerization of human Cu-Zn superoxide dismutase." Proc Natl Acad Sci U S A **104**(24): 10040-10044.

Stryer, L. (1988). Biochemistry. New York, USA, W.H. Freeman and Company.
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): W610-W614.
Teilum, K., J. Olsen, et al. (2009a). "Functional aspects of protein flexibility." Cell Mol Life Sci **66**(14): 2231-2247.
Teilum, K., M. H. Smith, et al. (2009b). "Transient structural distortion of metal-free Cu/Zn superoxide dismutase triggers aberrant oligomerization." Proc Natl Acad Sci U S A **106**(43): 18273-18278.
Tirion, M. M. (1996). "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis." Physical Review Letters **77**(9): 1905-1908.
Tjandra, N., S. E. Feller, et al. (1995). "Rotational diffusion anisotropy of human ubiquitin from 15N NMR relaxation." J Am Chem Soc **117**: 12562-12566.
Tjandra, N., J. G. Omichinski, et al. (1997). "Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution." Nat Struct Mol Biol **4**: 732-738.
Ulmschneider, J. and W. Jorgensen (2003). "Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias." J Chem Phys **118**: 4261-4272.
Ulrich, E. L., H. Akutsu, et al. (2007). "BioMagResBank." Nucleic Acids Res **36**: D402-D408.
Uppenberg, J., N. Ohrner, et al. (1995). "Crystallographic and molecular-modeling studies of lipase B from Candida antarctica reveal a stereospecificity pocket for secondary alcohols." Biochemistry **34**: 16838-16851.
van Gunsteren, W. F., S. R. Billeter, et al. (1996). Biomolecular Simulation: The GROMOS96 Manual and User Guide, vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.
Vriend, G. (1990). "What If - a molecular modeling and drug design program." J Mol Graph **8**(1): 52-&.
Watt, E. D., H. Shimada, et al. (2007). "The mechanism of rate-limiting motions in enzyme function." Proc Natl Acad Sci U S A **104**(29): 11981-11986.
Wells, S., S. Menor, et al. (2005). "Constrained geometric simulation of diffusive motion in proteins." Phys Biol **2**(4): 127-136.
Wlodarski, T. and B. Zagrovic (2009). "Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin." Proc Natl Acad Sci U S A **106**(46): 19346-19351.
Wlodawer, A., L. A. Svensson, et al. (1988). "Structure of phosphate-free ribonuclease A refined at 1.26Å." Biochemistry **27**: 2705-2717.
Yang, L., G. Song, et al. (2009). "Protein elastic network models and the ranges of cooperativity." Proc Natl Acad Sci U S A **106**(30): 12347-12352.
Zachariae, U., R. Schneider, et al. (2008). "The molecular mechanism of toxin-induced conformational changes in a potassium channel: relation to C-type inactivation." Structure **16**: 742-754.
Zheng, W., B. R. Brooks, et al. (2007). "Protein conformational transitions explored by mixed elastic network models." Proteins **69**(1): 43-57.

2.4 Generative Probabilistic Models Extend the Scope of Inferential Structure Determination

This manuscript describes a successful combination of probabilistic models and MC simulations for protein structure determination. The approach is based on the idea of conducting a Monte Carlo simulation using probabilistic models of local structure as trial distributions, and modeling the non-local interactions using an energy function derived from Nuclear Overhauser Effect measurements. In this study, CRISP moves are combined with standard MC methods for sampling the conformational space. The method is tested on two model systems, and the results show improved efficiency and accuracy compared to the current state-of-the-art technique.

This is a co-author article, I was involved in the design and implementation of the method used to combine the MC moves together with the probabilistic model of local structure, and I contributed to the optimization of the move-set for this specific application.

Journal of Magnetic Resonance 213 (2011) 182–186

Contents lists available at SciVerse ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Communication

Generative probabilistic models extend the scope of inferential structure determination

Simon Olsson [a], Wouter Boomsma [b], Jes Frellsen [a], Sandro Bottaro [b], Tim Harder [a], Jesper Ferkinghoff-Borg [b,*], Thomas Hamelryck [a,*]

[a] Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark
[b] Biomedical Engineering, DTU Elektro, Technical University of Denmark, Ørsteds Plads, DK-2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 20 June 2011
Revised 19 August 2011
Available online 6 September 2011

Keywords:
Inferential structure determination
Generative probabilistic models
Sparse data

ABSTRACT

Conventional methods for protein structure determination from NMR data rely on the *ad hoc* combination of physical forcefields and experimental data, along with heuristic determination of free parameters such as weight of experimental data relative to a physical forcefield. Recently, a theoretically rigorous approach was developed which treats structure determination as a problem of Bayesian inference. In this case, the forcefields are brought in as a prior distribution in the form of a Boltzmann factor. Due to high computational cost, the approach has been only sparsely applied in practice. Here, we demonstrate that the use of generative probabilistic models instead of physical forcefields in the Bayesian formalism is not only conceptually attractive, but also improves precision and efficiency. Our results open new vistas for the use of sophisticated probabilistic models of biomolecular structure in structure determination from experimental data.

1. Introduction

Current methods for macromolecular structure determination rely on the seminal idea of hybrid energy minimization introduced by Jack and Levitt [1]. However, the choice of model parameters, such as the weight of the experimental data with respect to a physical force field, is intrinsically problematic in this approach – a fact that was already recognized in the original study [1]. With a growing number of sources of experimental data used in protein structure determination, estimation of weights and other nuisance parameters is becoming increasingly problematic. Current methodology relies on a more or less arbitrary choice of these parameters, using heuristic approaches [2]. While a persistent concern towards the applied heuristics has been evident in the literature [3,2,4], only few quantitative methods have been described to rigorously determine these nuisance parameters [4,5]. These methods, and the underlying Bayesian approach are referred to as inferential structure determination (ISD).

Bayesian probabilistic inference has previously shown great potential in macromolecular structure determination [2,6]. However, the scope of the approach has been limited due to excessive computational demands. The current study describes a new approach to inferential structure determination which draws on the use of generative probabilistic models. Generative probabilistic models, or GPMs, are probabilistic models that allow sampling. Here, we demonstrate that the use of GPMs greatly increases efficiency, precision and scope of rigorous inferential structure determination. As these GPMs contain information about protein structure, they may supersede physical forcefields – especially in cases where data is very sparse.

2. Methods

In the ISD approach, samples are drawn from a joint posterior distribution over conformational space, X, and model parameter space, n, given experimental data, D, and prior knowledge, I:

$$p(X,n|D,I) \propto p(D|X,n,I)p(n|I)p(X|I).$$

Consequently, a natural result of posterior sampling is an ensemble of conformers representing the experimental uncertainty. That is, the Bayesian formalism accounts for uncertainty and degeneracy, a feature that is difficult to obtain when using schemes that minimize a hybrid energy consisting of a physical and a data-dependent term [7,8].

In ISD, a physical forcefield E_{phys} enters the Bayesian framework as a conformational prior through a canonical ensemble $p(X|I) \propto e^{-\beta E_{\text{phys}}}$, where $\beta = 1/kT$, k is Boltzmann's constant and T is the temperature [2]. The data enters as a likelihood function, $p(D|X,n,I)$; its product with the prior distributions, $p(n|I)p(X|I)$,

* Corresponding authors.
E-mail addresses: solsson@binf.ku.dk (S. Olsson), jfb@elektro.dtu.dk (J. Ferkinghoff-Borg), thamelry@binf.ku.dk (T. Hamelryck).

1090-7807/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2011.08.039

results in the posterior distribution, $p(X,n|D,I)$. When the posterior is defined in this way, Markov Chain Monte Carlo (MCMC) sampling requires evaluation of both likelihood and priors explicitly, in each step. This can potentially lead to substantial computational costs. Conversely, using no, or a uninformative forcefield, leaves a vast conformational space [9]. Here, we use GPMs of local protein structure instead of the Boltzmann distribution of a physical forcefield. Consequently, we demonstrate that the explicit evaluation of the prior can be avoided altogether as the information of the prior enters the posterior distribution through sampling.

Recently, our group has published several GPMs of protein conformational space, describing backbone (TorusDBN) [10,11] and sidechain (Basilisk) [12] dihedral angles. These models only provide structural information on a local sequential scale, ideally complementing the long-range information obtained from NMR nuclear Overhauser enhancements experiments (NOE). As generalizations of the commonly used fragment- [13] and rotamer-libraries [14], and related potentials that involve discretization [15], these GPMs also serve to reduce the complexity of the conformational space. The particular GPMs applied here use continuous angular probability distributions to avoid the intrinsic limitations caused by discretization [16]. Furthermore, since these GPMs are probability distributions, probabilities of arbitrary conformations can be evaluated, which is not generally possible for fragment- and rotamer libraries. Consequently, the full posterior probability can be evaluated explicitly when necessary. Here, we demonstrate that the use of GPMs as conformational proposal distributions can dramatically increase convergence in MCMC sampling of protein conformers from a posterior distribution, in addition to providing an increase in precision.

The GPMs, TorusDBN and Basilisk, enter the ISD approach as $p(X|I) \propto p(b|a)p(\chi|a)$, where a denotes amino acid sequence, while b and χ denote backbone and sidechain conformations respectively. Thus, during simulation we alternate between moving in backbone and sidechain conformational space, conditioned on amino acid sequence. Following Rieping et al. we assume idealized Engh–Huber bond lengths [17] and parameterize conformations as sets of torsion angles [18]. Variations in the bond angles were allowed to facilitate conformational sampling [19].

We used a generalized ensemble Metropolis–Hastings sampling scheme to draw samples from the posterior distribution. To prioritize search in relevant regions of the conformational space we adopted the $1/k$-ensemble implemented using the generalized multi-histogram equations [20,21]. The $1/k$-ensemble allows sampling independently of temperature, thus avoiding nuisance parameters such as the number of replicas, and their temperature span. It is, however, important to stress that the statistical information provided by this sampling scheme is equivalent to the Replica Exchange Monte Carlo scheme used in the original ISD study [22]. We employ the log-normal formulation of the NOE data to evaluate $p(D|b,\chi,n,I)$, as this provides the least biased formulation of the likelihood [5].

To assess the performance of TorusDBN and Basilisk as conformational priors, and for comparison to previous results, we created a set of conformers corresponding to the lowest posterior samples, using the very sparse (154 constraints) SH3 FYN domain data [2] and the TRP-Cage data set [28]. As a model baseline, we carried out the same simulations without the models of local protein structure. This simple hard-sphere potential corresponds to the use of a prior distribution reminiscent of that of the original ISD implementation [2].

2.1. Posterior sampling

As described previously, we sample from the joint posterior distribution $p(X,n|D,I)$ [2]:

$$p(X,\mathbf{n}|D,I) \propto \sigma^{-(n+1)}\gamma^{-1}\exp\left[-\frac{1}{2\sigma^2}\chi^2(d(\chi,b),I)\right]p(\chi|a)p(b|a),$$

with the log-normal chi-square: $\chi^2(d,D)=\sum_i^n \log^2(\gamma d_i^{-\alpha}/D_i), D_i$ are experimental data and d_i calculated distances [5]. χ and b are the sidechain and backbone dihedrals, respectively. γ and σ are ISPA (isolated spin-pair approximation) equilibration parameter and experimental uncertainty, respectively. A power $\alpha=-1$ was used here as all data were derived distances.

Here, we cannot employ the Gibbs sampling scheme applied in Rieping et al. [2], due to the inherent absence of an explicit temperature in the $1/k$ ensemble. This absence of an explicit temperature makes the implementation of the soft-sphere potential employed previously difficult without introduction of additional heuristics, and was therefore avoided [2]. Instead, we here use a Metropolis–Hastings approach, where the involved parameters are updated one at the time. The $1/k$ ensemble allows us to sample the conformational- and nuisance-space efficiently.

Low acceptance rates in the nuisance sampling was avoided by introducing a scheme exploiting the information about the current state. For the nuisance parameters, $n=\{\gamma,\sigma\}$, a log-change is proposed from a log-normal distribution with a standard deviation

$$\sigma_{n_i}=\frac{1.0}{\max\left(\left\|\frac{\partial \log p(X,n|D,I)}{\partial n_i}\right\|,1.0\right)}.$$

This expression was derived using standard error propagation and adds a simple regularizer which ensures a maximum standard deviation of 1.0 [23]. As a result, we can draw samples efficiently from the joint posterior distribution without the temperature dependent Gibbs sampling scheme. Using the log-normal distribution in this way we can ensure being in the right domain. We avoid additional bias from the log-normal distribution in the posterior, by dividing out the bias in the Monte Carlo acceptance ratio. For completeness, the analytical expressions of the standard deviations are shown here:

$$\sigma_\sigma=\frac{1.0}{\max\left(\left\|-\frac{\chi^2(d(\chi,b),I)}{\sigma^2}+v\right\|,1.0\right)}$$

where v is the number of datapoints, and:

$$\sigma_\gamma=\frac{1.0}{\max\left(\left\|\frac{v\log\gamma-\sum_i^v k_i}{\sigma^2}\right\|,1.0\right)},$$

with $k_i=\ln\frac{I_{\text{obs}}^i}{I_{\text{calc}}^i}$ corresponding to the log-ratio between the observed and back-calculated experimental data.

For sampling of the conformational space, a series of MCMC moves for backbone (pivot, local [19] and semi-local [24]) and sidechain conformations were employed. All applied moves fulfill detailed balance, and were chosen with even probability with respect to backbone and sidechain conformational and nuisance space. TorusDBN was extended to account for small deviations from ideal *cis*/*trans*-angles, using a normal distribution with mean at the ideal values and a standard deviation of five degrees. In the baseline model, all angles b, χ were sampled uniformly in the interval $[0,2\pi]$. Note that Basilisk was used in a backbone independent fashion for simplicity [12]. Samples were accepted or rejected according to the generalized $1/k$ ensemble [20]. Convergence was assessed through inspection of diagnostics provided by Muninn: the multi-histogram implementation of the generalized ensemble (http://www.muninn.sourceforge.net/). It is important to stress that convergence of histograms necessarily reflect convergence of posterior samples, additional sampling allow generation of more refined ensembles.

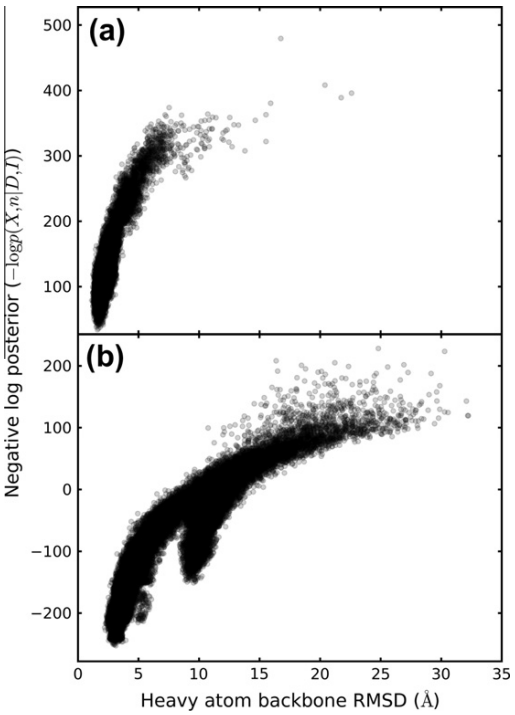


Fig. 1. Scatter plots of the RMSD of conformational samples to the crystal structure of SH3 FYN (PDB:1SHF chain A) versus $-\log p(X,n|D,I)$ (posterior density) for (a) TorusDBN and Basilisk and (b) the baseline prior after 400 million MCMC steps. Samples are from the $1/k$ ensemble.

Table 1
VADAR and PROCHECK structure quality statistics for the previously published ensemble (PDB: 1ZBJ) (**1ZBJ**) [2] and current SH3 FYN (**GPMs**) ensembles and reference values presented by VADAR (**Ref**). ϕ, ψ core, allowed, generous and outside denote distinct regions of the Ramachandran plot of decreasing favoredness. ω core denotes the percentage of ω-angles in the most favored region (the three other classes are not shown here). Packing defects, free energy folding, percentage of residues 95% buried and buried charges denotes the number of packing defects, free energy of folding and bury ratios for residues and charges, respectively [25]. Percentile reference values were normalized. PROCHECK *G*-factors reflect average log-odds of (ϕ,ψ), $(\chi1,\chi2)$, $(\chi1)$ and overall dihedral angle combinations.

VADAR Dihedral prior	**1ZBJ**	**GPMs**	**Ref**
ϕ, ψ core	68.95 ± 4.19%	88.33 ± 2.85%	91.84%
ϕ, ψ allowed	27.6 ± 4.12%	9.96 ± 3.17%	7.14%
ϕ, ψ generous	1.7 ± 1.27%	1.65 ± 1.50%	1.02%
ϕ, ψ outside	0.0 ± 0.0%	0.05 ± 0.0%	0.0%
ω core	100.0 ± 0.0%	91.0 ± 2.17%	97%
ω allowed	0.0 ± 0.0%	8.0 ± 2.61%	3%
ω generous	0.0 ± 0.0%	1.0 ± 1.49%	0%
Packing defects	11.95 ±2.85	5.95 ± 2.06	4.0
Free energy fold	−40.7 ± 1.88	−46.07 ± 2.06	−42.39
Res. 95% buried	2.25 ± 1.22	4.30 ± 1.90	6.0
Buried charges	0.15 ± 0.30	0.30 ± 0.56	0.0
PROCHECK Dihedral prior	**1ZBJ**	**GPMs**	
G-factor (ϕ,ψ)	−1.41	−0.72	
G-factor $(\chi1,\chi2)$	−1.82	0.25	
G-factor $(\chi1$ only)	−0.54	0.20	
G-factor (overall)	−1.43	−0.28	

3. Results

3.1. SH3 FYN

When employing the GPMs, the sampling of the posterior distribution defined by the sparse SH3 FYN data set converges in less than 36 h of computation time on a single standard CPU core. In comparison, the previously published ISD ensemble derived from the same data set took 3 days on a 50 core computer cluster [2]. Even given the increase in average computational power since 2005, this is a substantial increase in the efficiency. We do not observe convergence within the same simulation time when applying the baseline model. This illustrates clearly how the GPMs increase efficiency of posterior sampling.

Performing posterior sampling with the baseline prior, gives rise to two distinct conformational basins (Fig. 1b). There is an excited basin corresponding to the mirror image of the native basin. The local geometry of this basin is highly unfavorable. The second basin corresponds to the correct, native fold, observed in the crystal structure. The latter of the two basins is the only one observed when using the informative GPMs as conformational priors (Fig. 1a). Evidently, the experimental data likelihood in conjunction with the baseline prior only modestly distinguishes between the two folds, resulting in slow convergence due to an excessive conformational multiplicity. The basin with the correct fold is not thoroughly explored within the given time frame, resulting in relatively inaccurate structures among the 20 highest posterior conformer ensemble (Fig. 2b). In contrast, the ensembles obtained within the same simulation time using the TorusDBN and Basilisk priors accurately capture the native state (Fig. 2a). This result illustrates the importance of prior information to resolve degeneracies in sparse experimental data. While avoidance of poor

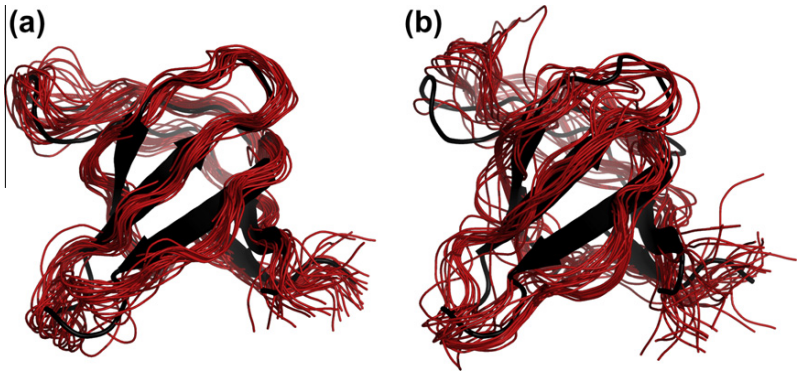


Fig. 2. Illustration of 20 of the samples with the highest posterior probability using (a) TorusDBN and Basilisk (RMSD: 1.74 ± 0.17Å) or (b) the baseline prior (RMSD: 3.12 ± 0.24 Å), after 400 million MCMC steps. Conformations are aligned to PDB: 1SHF chain A (shown in a black cartoon representation). Figure prepared using PyMOL (DeLano Scientific LLC).

stereochemistry has been pointed out previously as a feature of the ISD approach [2], degeneracy due to poor local structure has remained unaddressed.

The mean heavy-atom (C_α, C and N) root mean square deviation (RMSD) to the crystal structure from the 20 highest posterior probability structures (see Fig. 2) is comparable to the previously published ISD ensemble (1.84 ± 0.20 Å, PDB: 1ZBJ). However, statistics derived from structure validation server VADAR [25], WHATIF [26] and PROCHECK [27] were vastly improved (see Table 1 and Supplementary material) with respect to both packing quality and local structure. Importantly, clustering of (ϕ,ψ)-angle pairs in less favorable regions of the Ramachandran space is reduced dramatically (see SI). Other structure quality indicators such as number of buried charges remain unchanged. While the improvement in local

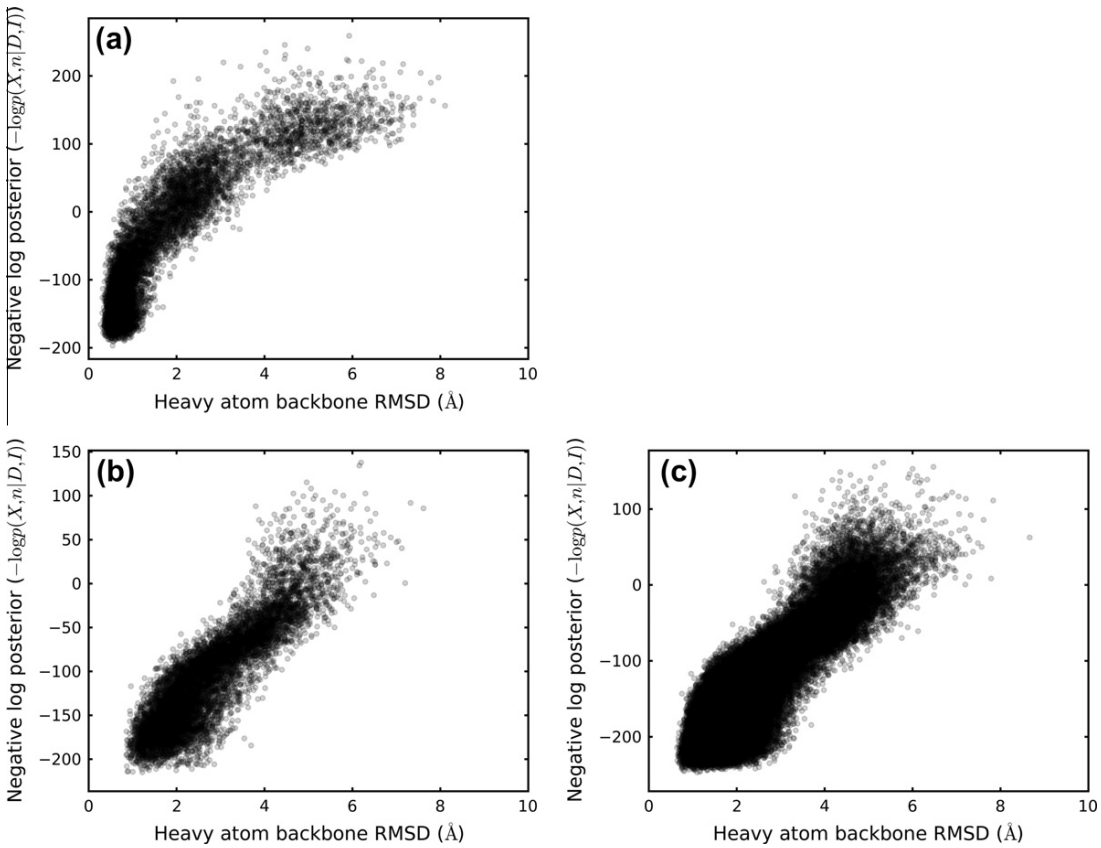


Fig. 3. Scatter plots of RMSD of conformational samples to the previously published NMR structure of TRP-Cage (PDB:1L2Y) versus $-\log p(X,n|D,I)$ (posterior density) for (a) TorusDBN and Basilisk and (b) baseline prior after 50 million MCMC steps; (c) baseline prior after 500 million MCMC steps. Samples are from the $1/k$ ensemble.

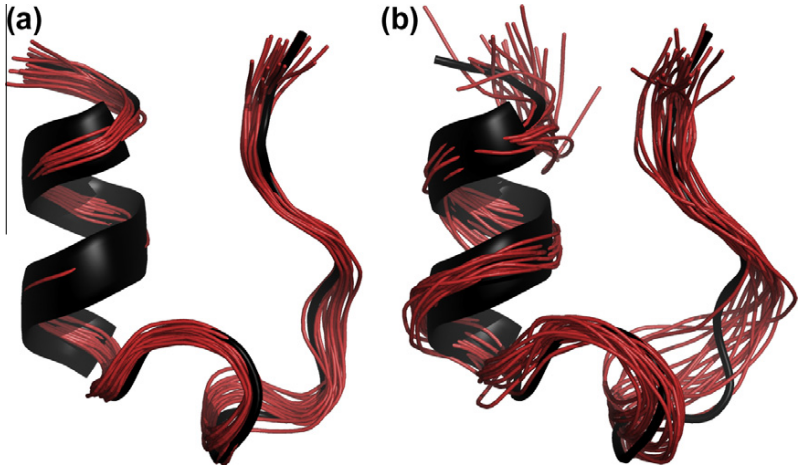


Fig. 4. Illustration of 20 of the samples with the highest posterior probability using (a) TorusDBN and Basilisk (RMSD: 0.63 ± 0.12 Å) or (b) baseline prior (RMSD: 1.41 ± 0.39 Å), after 50 million MCMC samples. Conformations are aligned to PDB: 1L2Y (shown in a black cartoon representation). Figure prepared using PyMOL (DeLano Scientific LLC).

structure is an expected consequence of the information contained in TorusDBN and Basilisk, non-local structure quality parameters such as packing defects hint increase in accuracy. The nuisance parameters σ, γ were estimated to be 0.11 ± 0.01 and 1.00 ± 0.01, respectively. These values deviate somewhat from the estimates obtained previously. The discrepancy may be linked to a different conformational prior distribution [2].

3.2. TRP-cage

In addition to increasing efficiency and precision, GPMs can account for the information derived from ambiguous NOE constraints. We demonstrate this point on the TRP-cage data set [28]. Of the reported 169 restraints, 37 involve pseudo atoms, which strictly speaking yields them ambiguous. In these particular calculations, the restraints were therefore not included. The resulting set of unambiguous NOE restraints are insufficiently informative to distinguish native-like structures from conformers with an RMSD of up to 3 Å from the previously published NMR structure. However, when we use the GPMs as structural priors, we obtain an ensemble of high resemblance with the previously published structure.

The simulations of TRP-cage were performed identically to those of SH3 FYN using 50 million MCMC steps. Both simulations complete within a few hours (see Fig. 3a and b). The pattern observed for SH3 FYN emerges again: when using GPMs convergence was reached within the simulation time, whereas convergence was not reached using the baseline model. Extending the simulation time with the baseline model to 500 million MCMC steps results in convergence (Fig. 3c). However, the resulting 20 highest posterior ensemble is of significantly lower quality (RMSD: 1.24 ± 0.39 Å) than the ensemble obtained using the GPMs running for 50 million MCMC steps, Fig. 4a. With these results we again demonstrate how efficiency is gained when employing GPMs in the ISD approach. In addition the results illustrate, how the unambiguous constraints [28] can be complemented by the local information contained in the GPMs.

4. Conclusions

In both examples presented here, the difference in accuracy of the selected ensembles is modest, with mean RMSD differences of at most 1 Å. However, the highest probability (or lowest energy) criterion for selection of conformation for these ensembles may not only underestimate the spread of the ensemble [29,30], but also ignore severe degeneracies (see Figs. 1 and 3). This points to the importance of using appropriate prior information when analyzing sparse data and suggests extra caution be taken when selecting these ensembles.

This communication describes how generative probabilistic models can be applied to significantly increase efficiency and precision of inferential structure determination. As a natural extension, we propose the development of more specialized GPMs, drawing on additional prior information such as protein family membership or chemical shifts. Such models would presumably resolve degeneracies to an even greater extent, further increasing the scope, efficiency and precision of the inferential structure determination approach.

Acknowledgments

We thank F.M. Poulsen, K. Lindorff-Larsen and J.H. Jensen for critically reading the manuscript. T. Harder is funded by the Danish Council for Strategic Research (NaBiIT, 2106-06-009). W. Boomsma is funded by the Danish Council for Independent Research (FNU, 272-08-0315). J. Frellsen and S. Olsson are funded by the Danish Council for Independent Research (FTP, 09-066546). S. Bottaro acknowledge funding from Radiometer, DTU.Elektro.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2011.08.039.

References

[1] A. Jack, M. Levitt, Refinement of large structures by simultaneous minimization of energy and R factor, Acta Crystallogr. A 34 (1978) 931–935.
[2] W. Rieping, M. Habeck, M. Nilges, Inferential structure determination, Science 309 (2005) 303–306.
[3] M. Williamson, C. Craven, Automated protein structure calculation from NMR data, J. Biomol. NMR 43 (2009) 131–143. doi:10.1007/s10858-008-9295-6.
[4] M. Habeck, W. Rieping, M. Nilges, Weighting of experimental evidence in macromolecular structure determination, Proc. Natl. Acad. Sci. USA 103 (2006) 1756–1761.
[5] W. Rieping, M. Habeck, M. Nilges, Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures, J. Am. Chem. Soc. 127 (2005) 16026–16027.
[6] C.K. Fisher, A. Huang, C.M. Stultz, Modeling intrinsically disordered proteins with bayesian statistics, J. Am. Chem. Soc. 132 (2010) 14919–14927.
[7] C. Schwieters, J. Kuszewski, N. Tjandra, G. Clore, The Xplor-NIH NMR molecular structure determination package, J. Magn. Reson. 160 (2003) 66–74.
[8] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T. Malliavin, M. Nilges, Aria2: automated NOE assignment and data integration in NMR structure calculation, Bioinformatics 23 (2007) 381–382.
[9] M. Habeck, Statistical mechanics analysis of sparse data, J. Struct. Biol. (2010).
[10] T. Hamelryck, J.T. Kent, A. Krogh, Sampling realistic protein conformations using local structural bias, PLoS Comput. Biol. 2 (2006) e131.
[11] W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, T. Hamelryck, A generative, probabilistic model of local protein structure, Proc. Natl. Acad. Sci. USA 105 (2008) 8932–8937.
[12] T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K.E. Johansson, T. Hamelryck, Beyond rotamers: a generative, probabilistic model of side chains in proteins, BMC Bioinf. 11 (2010) 306.
[13] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, J. Mol. Biol. 268 (1997) 209–225.
[14] S.C. Lovell, J.M. Word, J.S. Richardson, D.C. Richardson, The penultimate rotamer library, Proteins 40 (2000) 389–408.
[15] J. Kuszewski, A.M. Gronenborn, G.M. Clore, Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases, Prot. Sci. 5 (1996) 1067–1080.
[16] G.L. Butterfoss, B. Kuhlman, Computer-based design of novel protein structures, Annu. Rev. Biophys. Biomol. Struct. 35 (2006) 49–65.
[17] R.A. Engh, R. Huber, Accurate bond and angle parameters for X-ray protein structure refinement, Acta Cryst. A47 (1991) 392–400.
[18] L.M. Rice, A.T. Brünger, Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement, Proteins 19 (1994) 277–290.
[19] S. Bottaro, W. Boomsma, K.E. Johansson, C. Andreetta, T. Hamelryck, J. Ferkinghoff-Borg, Realizing the potential of Monte Carlo: subtle kinetics in dense protein systems, unpublished results.
[20] B. Hesselbo, R.B. Stinchcombe, Monte Carlo simulation and global optimization without parameters, Phys. Rev. Lett. 74 (1995) 2151–2155.
[21] J. Ferkinghoff-Borg, Optimized Monte Carlo analysis for generalized ensembles, Eur. Phys. J. B 29 (2002) 481–484.
[22] M. Habeck, M. Nilges, W. Rieping, Replica-exchange Monte Carlo scheme for Bayesian data analysis, Phys. Rev. Lett. 94 (2005) 018105-1–018105-4.
[23] S. Meyer, Data Analysis for Scientists and Engineers, Wiley, 1975.
[24] G. Favrin, A. Irbäck, F. Sjunnesson, Monte Carlo update for chain molecules: biased Gaussian steps in torsional space, J. Chem. Phys. 114 (2001) 8154–8158.
[25] L. Willard, A. Ranjan, H. Zhang, H. Monzavi, R.F. Boyko, B.D. Sykes, D.S. Wishart, Vadar: a web server for quantitative evaluation of protein structure quality, Nucleic Acids Res. 31 (2003) 3316–3319.
[26] G. Vriend, WHAT IF: a molecular modeling and drug design program, J. Mol. Graph. 8 (1990). 52–56, 29.
[27] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, Procheck: a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283–291.
[28] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Designing a 20-residue protein, Nat. Struct. Biol. 9 (2002) 425–430.
[29] D. Zhao, O. Jardetzky, An assessment of the precision and accuracy of protein structures determined by NMR: dependence on distance errors, J. Mol. Biol. 239 (1994) 601–607.
[30] C.A. Spronk, S.B. Nabuurs, A.M. Bonvin, E. Krieger, G.W. Vuister, G. Vriend, The precision of NMR structure ensembles revisited, J. Biomol. NMR 25 (2003) 225–234. 10.1023/A:1022819716110.

2.5 Implicit Solvent Model Parameterization Via Relative Entropy Minimization

In this article we derive a parameterization for a widely used implicit solvent model from atomistic state-of-the-art molecular dynamics simulations. A recently-proposed coarse-graining technique based upon the minimization of the relative entropy between the coarse-grained and the all-atom ensembles is used. We present preliminary results, and discuss the validity of the approach.

This is a first-author article, and I was involved in all aspects of the work. The project was carried out under the supervision of Dr. Robert Best at the Department of Chemistry, University of Cambridge (United Kingdom) and Prof. Kresten Lindorff-Larsen, at the Department of Biology, Copenhagen University (Denmark).

Implicit solvent model parameterization via relative entropy minimization

Sandro Bottaro [1], Kresten Lindorff-Larsen[2], Robert B. Best [3]

1 Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

2 Department of Biology, University of Copenhagen, Copenhagen, Denmark

3 Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Abstract

An efficient solvation term based on a Gaussian solvent-exclusion model is combined with the all-atom CHARMM 36 force-field for simulations of proteins in aqueous environment. The presented model is obtained by extending to full atomic detail the EEF1 effective energy function, which is based on the united-atom representation of the CHARMM 19 force-field. The model parameters are adjusted so as to reproduce the behavior of explicit water simulations. To reach this goal, a recently proposed coarse-graining method based on the iterative minimization of an entropy-related objective function is used. The procedure is first validated and subsequently employed to produce a set of optimized model parameters using the α-helical $(AAQAA)_3$ peptide as a model system. The resulting effective energy function, termed EEF1-SB, is subjected to a number of tests. Molecular dynamics simulations at room temperature of two proteins in their native conformation are performed, and stable trajectories are obtained, showing improved or similar accuracy compared to existing methodologies. The range of applicability of EEF1-SB is further assessed by performing folding simulations on structured peptides. The obtained results show that EEF1-SB correctly folds β-stranded peptides, but fails to detect the structural propensity of α-helical systems. EEF1-SB thus provides an efficient (only 20% slower compared to *vacuum* simulations) and realistic first approximation for treating solvent effects. However, the observed propensity towards β conformations suggests that a further optimization of the force-field is needed.

Introduction

The aqueous environment plays an important role in determining the function, structure and dynamics of biomolecules [1,2]. For this reason, the solute-water interaction has been the subject of many theoretical, computational and experimental studies over the last decades [3–5].

The most realistic way to treat solvation effects in computer simulations is the inclusion of explicit water molecules. The high level of detail provided by this approach, however, has substantial computational costs, which often become prohibitive for molecular systems undergoing significant structural rearrangements.

Implicit solvent represents a simplified, though less accurate, alternative to explicit water models for treating solvent effects. In implicit solvent models the average influence of water molecules is described by an effective energy that depends only on the atomic coordinates of the solute. The formulation of an accurate and computationally efficient description of solvent effects is a nontrivial theoretical problem, and the development of implicit water models for biomolecular simulations has progressed along many different lines of research. One of the most simple approaches is given by solvent-accessible surface area (SASA) models [6], in which the solvent effect is taken proportional to the area of the solute atom that is accessible to solvent molecules. Typically, the proportionality constants are determined by matching the experimental free energy of hydration of small molecules [6,7] or by reproducing the solvent accessible surface area for a selected set of systems [8,9]. Similar approaches have been employed to parameterize contact models, where the solvation free energy depends on the number of contacts that each atom makes with other solute particles [10], and the contacts are weighted according to some function of their distance [11]. Another alternative approach is represented by Gaussian solvent-exclusion models, in which the solvent effects are assumed to be proportional to the volume of the first hydration shell that is accessible to the solvent [12]. Because of its good accuracy and efficiency (only about 50% more computational effort with respect to a *vacuum* simulation), the Gaussian solvent-exclusion model EEF1 [13] was applied to a wide range of biological problems. Despite known limitations [14–16], EEF1 provides a reasonably accurate description of solvent

effects [13,17,18], it has been shown in certain cases to yield comparable results with respect to explicit water simulations [19], and recently led to successful applications in protein structure prediction [20,21] and folding studies [22].

In all of the aforementioned effective potentials, the electrostatic effects are usually crudely approximated (e.g. using a distance-dependent dielectric constant [9,13]) or completely ignored. The effect of the solvent on electrostatic interactions is more rigorously described in continuum electrostatic models, where the solute is assumed as a low-dielectric cavity immersed in a high-dielectric and featureless environment. The electrostatic (polar) free energy of solvating a molecule is then calculated solving the Poisson-Boltzmann (PB) equation [23–25] or efficiently estimated by using the popular Generalized-Born (GB) equation [26,27]. A number of variations to the original GB approach have been introduced in order to improve the accuracy and the efficiency of the method [28–32]. While continuum electrostatic models provide a theoretical formulation for polar interactions, the non-polar effects (e.g. hydrophobicity) are usually neglected or modeled by empirical potentials. Moreover, it is worth highlighting that the computational complexity of most PB and GB methods scales very poorly with the system size (although notable exceptions exist [31,32]) and is comparable to explicit water simulations for large, globular molecules [33].

In the present work we extend the effective energy function EEF1 [13], which was originally based on the united-atom CHARMM 19 force-field [34], to the all-atom description of CHARMM 36 [35]. The significant differences in the molecular representation as well as in the parameterization of the two force-fields does not allow a direct transfer of the EEF1 model parameters. Therefore, we devised a modified version of EEF1, which we term EEF1-SB, were the model parameters are adjusted so as to mimic the equilibrium ensemble obtained from explicit water simulations. Following the idea of Shell and co-workers [36–39], the "overlap" between the implicit and explicit water models is maximized by using a procedure based on the iterative minimization of an objective function called relative entropy.

In the next sections we describe the EEF1-SB effective energy, and we briefly outline the

relative entropy minimization approach used for model parameterization. The method is first validated by showing that the explicit water structural ensemble of the short Ala_5 peptide is accurately reproduced by the coarse-grained model upon relative entropy minimization. Subsequently, the procedure is used to determine the optimal parameters in EEF1-SB, using the α-helical peptide $(AAQAA)_3$ as a model system. Finally, we test the accuracy of EEF1-SB by performing molecular dynamics simulations on the native state of globular proteins and by conducting folding studies on three short peptides.

The obtained results prove the EEF1-SB model to produce stable trajectories when simulating the near-native state of globular proteins, yielding better or similar results compared to existing empirical implicit solvent models. In the folding studies, EEF1-SB fails to reproduce the secondary structure propensities of α-helical systems, but correctly fold β-stranded peptides. It should be noted that similar biases are observed in many force-fields [40, 41], unless specific corrections are introduced. Overall, the results indicate that the effective energy function EEF1-SB developed here is a realistic but not fully accurate approximation for proteins in aqueous environment. Specifically, the clear propensity towards the β-regions of the Ramachandran space suggests that a further optimization of the force-field is desirable.

Description of the model

Solvent-exclusion model

The formulation of the solvent exclusion model in EEF1-SB is identical to the EEF1 approach [13], where the total solvation free energy of a protein is expressed as a sum of atomic contributions

$$\Delta G^{solv} = \sum_i \Delta G_i^{solv} \tag{1}$$

each individual term ΔG_i^{solv} is equal to a reference solvation free energy ΔG_i^{ref}, obtained by dissecting the experimental free energy for a set of small compounds into group contributions [42],

minus a reduction due to the presence of surrounding atoms

$$\Delta G_i^{solv} = \Delta G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j \tag{2}$$

Here, the sum runs over the neighboring atoms j with volume V_j, and the solvation free energy density $f_i(r)$ is a Gaussian function of the distance, chosen such that the volume integral over the first solvation shell of thickness λ accounts for $\approx 85\%$ of the solvation energy

$$f_i(r) 4\pi r^2 = \frac{2}{\sqrt{\pi}} \frac{\Delta G_i^{free}}{\lambda_i} \exp\left\{ -\frac{(r - R_i)^2}{\lambda_i^2} \right\} \tag{3}$$

where ΔG^{free} is the solvation free energy of the isolated group, and R the van der Waals radius.

Model parameters

While in the EEF1 effective energy the atom types are those used in the united-atom CHARMM 19 force-field [34], in which only hydrogen atoms belonging to polar groups are explicitly included, in the present work we employ an all-atom representation to be used in combination with the CHARMM 36 force-field [35].

The reference solvation free energies ΔG^{ref}, taken from the EEF1 model [13], were originally obtained by dissecting the experimental solvation free energy (at $T = 298.15K$) for a set of model compounds into group contributions [42], and subsequently corrected to account for long-range van der Waals effects [13]. A similar procedure was used to determine the solvation enthalpy ΔH [43] and heat capacity ΔC_p [44]. These quantities make it possible to obtain an approximate expression for ΔG^{ref} as a function of the temperature. In the development of EEF1-SB, the solvation free energy of the isolated atom ΔG^{free} is initially assumed to be equal to ΔG^{ref}, and then optimized by relative entropy minimization (see below).

The volume of each atom type is calculated as the van der Waals volume minus the overlap volume between covalently bonded atoms [29]. The CHARMM 36 van der Waals radii and bond lengths are used for the volume calculation. When the same atom type is found in different

covalent arrangements, the most common is used, and triple or higher overlap volumes are neglected. All hydrogens are assumed not to contribute to the solvation energy, and their volumes are set to zero.

Finally, the thickness of the hydration shell, λ, is set to the 3.5 Å except for the atoms in charged groups, for which a value of 6 Å is used. The final values for the different parameters are listed in table 1.

Treatment of non-bonded interactions and ionic groups

The electrostatic screening effect of water is not considered directly by the exclusion-solvent model, and is here approximated using a linear, distance-dependent dielectric constant (i.e. $\epsilon = r$). As pointed out by Lazaridis and Karplus [13], the distance-dependent dielectric constant does not screen the electrostatic interactions for charged groups to a sufficient degree. In order to account for this effect, ionic side-chains are neutralized by adjusting the partial atomic charges, as detailed in table 2. To model the strong interactions between atoms in ionic groups and solvent molecules, the correlation λ for these atom types is set to 6 Å, and the reference solvation free energies ΔG^{ref} are arbitrarily set to large values, in order to increase their hydrophilic propensity. Electrostatic and van der Waals interactions are smoothly switched-off between 7 and 9 Å, and interactions between atoms separated by three (1-4 pairs) covalent bonds are not rescaled.

Atom type	Volume	ΔG^{ref}	ΔG^{free}	ΔH	ΔCp	λ
C	14.720	0.000	0.000	0.000	0.0	3.5
CD	14.720	0.000	0.000	0.000	0.0	3.5
CT1	11.507	-0.187	-0.160	0.876	0.0	3.5
CT2	18.850	0.372	0.318	-0.610	18.6	3.5
CT2A	18.666	0.372	0.318	-0.610	18.6	3.5
CT3	27.941	1.089	0.930	-1.779	35.6	3.5
CPH1	5.275	0.057	0.068	-0.973	6.9	3.5
CPH2	11.796	0.057	0.068	-0.973	6.9	3.5
CPT	4.669	-0.890	-0.760	2.220	6.9	3.5
CY	10.507	-0.890	-0.760	2.220	6.9	3.5
CP1	25.458	-0.187	-0.160	0.876	0.0	3.5
CP2	19.880	0.372	0.318	-0.610	18.6	3.5
CP3	26.731	0.372	0.318	-0.610	18.6	3.5
CC	16.539	0.000	0.000	0.000	0.0	3.5
CAI	18.249	0.057	0.049	-0.973	6.9	3.5
CA	18.249	0.057	0.049	-0.973	6.9	3.5
N	0.000	-1.000	-0.854	-1.250	8.8	3.5
NR1	15.273	-5.950	-5.081	-9.059	-8.8	3.5
NR2	15.111	-3.820	-3.262	-4.654	-8.8	3.5
NR3	15.071	-5.950	-5.081	-9.059	-8.8	3.5
NH1	10.197	-5.950	-5.081	-9.059	-8.8	3.5
NH2	18.182	-5.950	-5.081	-9.059	-8.8	3.5
NH3	18.817	-20.000	-17.078	-25.000	-18.0	6.0
NC2	18.215	-10.000	-8.539	-12.000	-7.0	6.0
NY	12.001	-5.950	-5.08	-9.059	-8.8	3.5
NP	4.993	-20.000	-17.078	-25.000	-18.0	6.0
O	11.772	-5.330	-4.551	-5.787	-8.8	3.5
OB	11.694	-5.330	-4.551	-5.787	-8.8	3.5
OC	12.003	-10.000	-8.539	-12.000	-9.4	6.0
OH1	15.528	-5.920	-5.055	-9.264	-11.2	3.5
OS	6.774	-2.900	-2.476	-3.150	-4.8	3.5
S	20.703	-3.240	-2.767	-4.475	-39.9	3.5
SM	21.306	-3.240	-2.767	-4.475	-39.9	3.5

Table 1. Solvation parameters used in EEF1-SB. The volume (in $Å^3$) is given by the van der Waals volume minus the overlap with the volume of covalently bonded atoms. The values of ΔG^{ref} (kcal mol^{-1}), ΔH (kcal mol^{-1}), and ΔCp (kcal mol^{-1} K^{-1}) are taken from the original EEF1 model [13]. ΔG^{free} (in kcal mol^{-1}) were optimized using the relative entropy minimization procedure. The correlation length λ is 3.5 Å , except for atoms in ionic groups, for which a value of $\lambda = 6$ Å is used.

Residue	Atom	Charge	Residue	Atom	Charge
ARG	CD	-0.30	HSP	CB	-0.10
	HD1/HD2	0.05		HB1/HB2	0.05
	NE	-0.28		CD2	0.05
	HE	0.12		HD2	0.00
	CZ	-0.20		CG	0.05
	NH1/NH2	-0.121		NE2/ND1	-0.55
	HH1/HH2	0.2005		HE2/HD1	0.45
				CE1	0.10
ASP	CB	-0.28		HE1	0.00
	HB1/HB2	0.14			
	HB1/HB2	0.14	LYS	CE	0.00
	HB1/HB2	0.14		HE1/HE2	0.00
				NZ	-0.90
GLU	CG	-0.28		HZ1/HZ2/HZ3	0.30
	HG1/HG2	0.14			
	CD	1.00			
	OE1/OE2	-0.50	NTER	N	-0.90
				HT1/HT2/HT3	0.20
GLP	CG	-0.21		HA	0.10
	HG1/HG2	0.09		CA	0.20
	CD	0.75			
	OE1	-0.55	CTER	C	1.00
	OE2	-0.61		OT1/OT2	-0.50
	HE2	0.44			

Table 2. Partial atomic charges for ionic groups in EEF1-SB

Relative entropy minimization

ΔG^{free} optimization

The optimal ΔG^{free} were determined using the relative entropy approach introduced by Shell and co-workers [36] as a general coarse-graining technique. The approach makes it possible to optimize the parameters in a coarse-grained potential, here represented by the the implicit solvent model in EEF1-SB, in order to reproduce the properties of a reference, all-atom potential. In the present work, the all-atom model is given by the CHARMM 36 force-field in combination with the TIP3P explicit water model [45].

The approach is based upon variational minimization of the relative entropy, S_{rel}, between the configurational ensembles produced by a reference explicit water (E) and implicit water (I) simulation

$$S_{rel} = \sum_i p_E(i) \log \frac{p_E(i)}{p_I(i)} \tag{4}$$

$p(i)$ is the probability of configuration i in the ensemble, and the index i proceeds over the all-atom configurations. In the canonical ensemble, the relative entropy is given by

$$S_{rel} = \beta(< U_I - U_E >_E) - \beta(A_I - A_E) + < S_{map} >_E \tag{5}$$

Here, U_E, U_I are the all-atom and coarse-grained potentials, respectively, $A = -k_B T log Z$, where Z is the partition function and $\beta = 1/k_B T$ is the inverse temperature T multiplied by the Boltzmann's constant k_B. The mapping entropy $< S_{map} >_E$ is the average entropy that results from degeneracies in the target-model mapping. According to Eq. 5, calculating S_{rel} requires the impractical estimation of free energies. Assuming the coarse-grained potential to be function of some parameters η, however, the derivatives of the relative entropy with respect to η can be

expressed as simple averages over the two ensembles

$$\begin{aligned}\frac{\partial S_{rel}}{\partial \eta} &= \beta\langle\frac{\partial U_I}{\partial \eta}\rangle_E - \beta\langle\frac{\partial U_I}{\partial \eta}\rangle_I \\ \frac{\partial^2 S_{rel}}{\partial \eta^2} &= \beta\langle\frac{\partial^2 U_I}{\partial \eta^2}\rangle_E - \beta\langle\frac{\partial^2 U_I}{\partial \eta^2}\rangle_I + \beta^2\langle\frac{\partial U_I}{\partial \eta}^2\rangle_I - \beta^2\langle\frac{\partial U_I}{\partial \eta}\rangle_I^2 \end{aligned} \tag{6}$$

Hence, standard numerical techniques can be employed to minimize the relative entropy with respect to the model parameters, for example by iterative application of the Newton-Raphson update rule

$$\eta_{k+1} = \eta_k - \gamma \left[\frac{\partial^2 S_{rel}}{\partial \eta^2}\right]^{-1} \left[\frac{\partial S_{rel}}{\partial \eta}\right] \tag{7}$$

Performing the minimization can be challenging in practical applications. First, because the absolute value of the relative entropy in Eq. 5 cannot be easily calculated. As proposed by Shell and co-workers [38], an approximate expression for the relative entropy is obtained from Eq. 5 via standard free energy perturbation [46]

$$S_{rel} \approx \log\left\{\langle \exp\left[\Delta - \langle\Delta\rangle_E\right]\rangle_E\right\} \tag{8}$$

where $\Delta = \beta(U_I - U_E)$. However, the approximation holds as long as a substantial overlap exists between the coarse-grained and all-atom ensembles. Moreover, the average of the exponential in Eq. 8 is dominated by individual contributions with large Δ, and can therefore be affected by large statistical errors. It is worth highlighting that for parameters linear in the potential U_I, the second derivative of the relative entropy in Eq. 6 reads

$$\frac{\partial^2 S_{rel}}{\partial \eta^2} = \beta^2(\langle\frac{\partial U_I}{\partial \eta}^2\rangle_I - \langle\frac{\partial U_I}{\partial \eta}\rangle_I^2) \tag{9}$$

This quantity is positive definite, and requiring the gradient $|\frac{\partial S_{rel}}{\partial \eta}|$ to be zero is a sufficient condition for optimality. As a consequence, for parameters linear in the potential, it is not strictly necessary to monitor the absolute value of the relative entropy during the minimization

procedure. Another difficulty is given by the fact that the calculation of averages over the coarse-grained ensembles (Eq. 6) requires a substantial computational effort, as a new implicit solvent simulation (with a new set of parameters η) has to be performed at each iteration step, although re-weighting techniques can in principle be used [47].

Finally, the minimization procedure becomes unstable (and inaccurate) if the individual implicit solvent simulations are not sufficiently equilibrated, causing for example large fluctuations during successive iteration of parameter optimization. In order to alleviate this problem, in the present work we make extensive use of replica exchange molecular dynamics (REMD) simulations [48]. The numerical stability of the relative entropy minimization is further improved by adjusting the step size in parameter space (i.e. the value of γ in Eq. 7). Following the ideas of Shell and co-workers [39], we dynamically adjust the step size until two heuristic criteria are met: i) the absolute change in the parameter at each iteration step is smaller than the 50% of its initial value and ii) the change $\Delta S_{rel} = S_{rel}(\eta_{k+1}) - S_{rel}(\eta_k)$ is smaller than the 20% of $S_{rel}(\eta_k)$, where the variation ΔS_{rel} is extimated via Zwanzig perturbation as

$$\Delta S_{rel} = \log\left(\langle \exp\left(-\beta \Delta U\right)\rangle\right) + \beta\langle \Delta U\rangle \tag{10}$$

where $\Delta U = U(\eta_{k+1}) - U(\eta_k)$, and the averages are computed over the η_k ensemble.

As formulated here, the relative entropy approach can be easily extended to perform a simultaneous optimization of multiple parameters. Therefore, it is in principle possible to adjust all the parameters of the implicit solvent model in EEF1-SB (i.e. the volume V, the reference free energy ΔG^{free} and the correlation length λ). However, the optimization in a high-dimensional space can be numerically unstable and leads to over-fitting when multiple atom-types are considered. For this reason, in the present work we used a minimal number of adjustable parameters. During the developmental stage we included both ΔG^{free} and λ in the minimization procedure, but the values of the latter parameter were subject to small variations during optimization. We therefore kept λ, as well as V, fixed to the values listed in table 1.

Relative entropy minimization on Ala_5

As a first test, we optimized the values of ΔG^{free} in order to reproduce the behavior of a 100ns MD simulation in explicit water of the short Ala_5 peptide. The results obtained on this system are useful to understand some of the difficulties of the approach, arising not only from the numerical and computational issues connected to the minimization, but also from the inability of the coarse-grained model to reproduce the effect of explicit water molecules.

In this example, we considered only the solvation parameters ΔG^{free} associated to the CA, CB, O and N backbone atoms as adjustable model parameters, excluding the atom-types of the N and protonated C terminus from the relative entropy minimization. At each step of the iterative procedure, a 5ns REMD simulation (8 replicas spanning the temperature range 298-650K) in implicit water was performed, using only the last 2.5ns for the actual calculation of the derivatives. The model parameters were updated using the standard Newton-Raphson method (Eq. 7), and 100 iteration steps were sufficient to observe the convergence of the procedure (Fig. 1a).

Although not to a perfect level of detail, the parameterization obtained via relative entropy minimization improves the similarity of the implicit solvent structural ensemble with respect to the explicit solvent equivalent. This is noteworthy, because the method aims to match the energetic distribution in the coarse-grained and all-atom system, and does not directly consider structural parameters. Fig. 1b shows the radius of gyration distribution for the explicit water simulation (black), for the implicit model with initial parameters (green) and with the optimized values of ΔG^{free} (pink). It is interesting to observe that the distribution for the optimized parameters is strikingly similar to the explicit water simulation, except for the shoulder of the distribution around 4 Å. As Fig. 1c suggests, this discrepancy is due to the fact that the optimized implicit solvent model does not sufficiently populate the states associated to the α-helical and polyproline II (PPII) region of the Ramachandran map. It should also be noted that the final set of ΔG^{free} are considerably different from the initial parameters (Fig. 1a), and in some cases (i.e. atom-types CB and O), the minimization procedure causes the sign inversion of ΔG^{free}. This

undesirable behavior is due to the fact that not all the solvent effects can be captured by the implicit solvent model. Specifically, the neutralization of the charged N terminus, together with the distance-dependent dielectric constant, poorly mimic the electrostatic interactions, which are likely to play a large role on this system. Therefore, large changes in the parameters are introduced by the relative entropy optimization, in order to compensate for effects not pertaining to the solvent-exclusion model itself. We also observe that a fairly large number of iteration steps (100) were performed to ensure convergence. Minimizing the relative entropy for a larger system not only requires longer simulations at each iteration step, but also implies that a search in a higher-dimensional parameter space has to be performed (i.e. the number of atom-types), thus considerably increasing the computational cost of the procedure.

Relative Entropy minimization on $AAQAA_3$

While the test on Ala5 is instructive and useful to validate the relative entropy approach, we use the Ac-$(AAQAA)_3$-NH_2 (which we will refer to as $(AAQAA)_3$) peptide as a model system to obtain optimized parameters in EEF1-SB. $(AAQAA)_3$ is a weakly structured peptide that significantly populates helical states under physiological conditions [49]. This characteristic secondary structure propensity, together with its small size, makes of $(AAQAA)_3$ an ideal model system for force-fields parameterization [40, 50, 51]. Notably, the CHARMM 36 force-field was optimized to correctly reproduce the behavior of $(AAQAA)_3$ in solution as inferred from chemical shift data, suggesting that the equilibrium ensemble obtained from explicit water simulations on this system can be reliably used as target distribution in the relative entropy minimization.

With the purpose of generating an equilibrated ensemble, we conducted a 150ns REMD simulation (32 replicas spanning a temperature range from 278 to 416 K) on $(AAQAA)_3$ in explicit water, and used the last 50ns of the trajectory as the target distribution. At each step of the relative entropy minimization, a 25ns REMD simulation in implicit solvent (16 replicas spanning the temperature range 285-570K) was performed, and only the last 12.5ns were used for calculating the Newton-Raphson update. In order to reduce the dimensionality

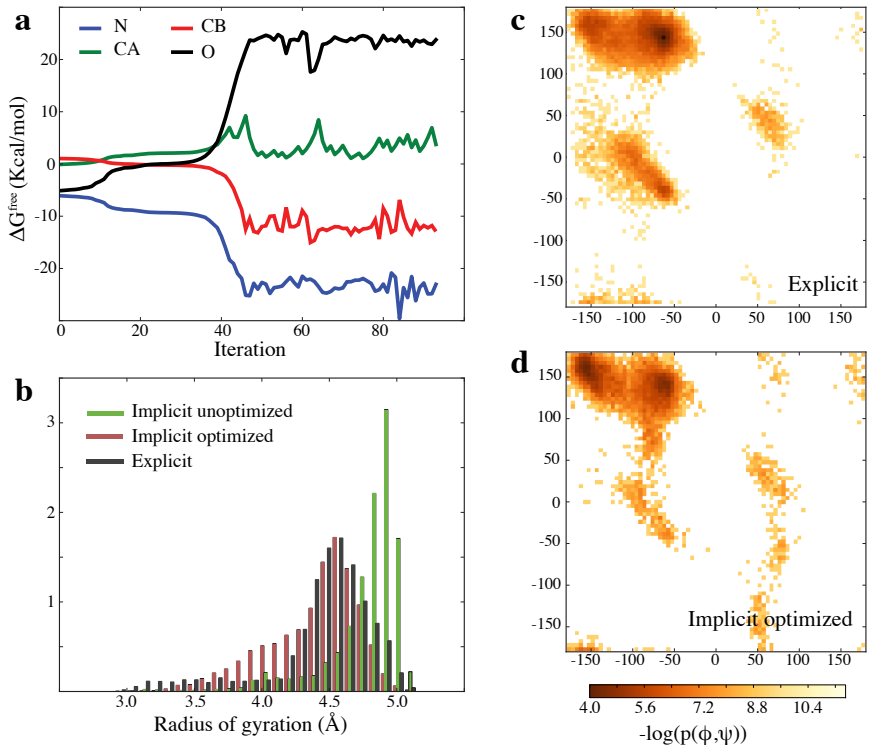


Figure 1. Relative entropy minimization on Ala_5. (**a**) Convergence behavior of the four parameters involved in the relative entropy minimization on Ala_5. (**b**) Radius of gyration distribution calculated on the equilibrium ensemble of the explicit solvent and implicit solvent before/after the optimization.(**c-d**) Ramachandran map $(-\log p(\phi, \psi))$ for the 3 central residues of the explicit and implicit solvent simulations.

of the problem and to avoid "runaway" of the parameters, we optimize a global scaling factor k for all ΔG^{free}, instead of considering one parameter for each atom-type. Starting from the initial value $k = 1.0$, the procedure converged in 20 iteration steps, yielding a global scaling factor $k = 0.854$ at optimality (i.e $\Delta G^{free}_{optimal} = k\Delta G^{free}_{initial} = k\Delta G^{ref}$ for all atom-types).

We assessed the effect of the relative entropy minimization by comparing 500ns REMD simulations (16 replicas spanning a 278-525 K temperature interval) using both the optimized and initial (unoptimized) parameter set. As shown in Fig. 2a, the changes in the implicit model parameters weakly affect the radius of gyration distribution at T=298K, that exhibits in both cases a pronounced peak around 8 Å. Specifically, the implicit solvent model favors hairpin-like structures, corresponding to the highly populated β region of the Ramachandran maps shown in Fig. 2b, region 3. Conversely, the broad distribution observed in the explicit water simulation reflects the structural diversity of the equilibrium ensemble, composed by a mixture of elongated polyproline II (PPII) and α-helical configurations (Fig. 2b, regions 1-2). According to experimental data [49], a significant helical content (≈ 0.21 at 300K) is observed in explicit water and, to a lesser extent, also for the optimized implicit solvent, while is completely absent in the unoptimized model (Fig. 2c). It is also interesting to observe that when helix is not formed in implicit solvent, turn or hairpin structures are obtained, suggesting that the solvent model does not stabilizes to a sufficient degree the unstructured, solvent-exposed states associated with the PPII region of the Ramachandran map (Fig. 2b, region 1). As described in a number of experimental studies [52, 53], these PPII conformations are stabilized by direct interactions between the main chain and water molecules, that are difficult to describe with a simple solvent-exclusion model.

Given the modest results achieved by minimizing the relative entropy on $(AAQAA)_3$, a number of attempts were made to improve the effectiveness of procedure. More precisely, we independently optimized ΔG^{free} (as in the validation test on Ala_5), or included all backbone atom-types and the correlation length λ as adjustable model parameter. Unfortunately, none of the above approaches produced a better agreement with the explicit solvent ensemble. The

current study thus serves to illustrate the improvements that can be obtained by optimizing only the implicit solvent model and within the context of the current EEF1 functional form.

Model Validation

Simulations of native proteins

We first show EEF1-SB to yield stable trajectories by performing molecular dynamics simulations at room temperature for two folded proteins, ubiquitin (76 residues, pdb code 1UBQ) and GB3 (56 residues, pdb code 1P7E). As a number of experimental and computational studies suggests, both systems are very stable, but at the same time undergo small conformational changes occurring on the microsecond time-scale [54, 55]. In the present test, we simply assess the ability of the solvent model to maintain a native-like structure during the simulation, and to avoid the main problems that arise in *in vacuo* simulations (e.g. compactification and large deviation from the native conformation). Starting from the experimentally solved structures, we performed 100ns molecular dynamics simulations at T=300K using the EEF1-SB model. Both systems remained in the vicinity of the native structure throughout the simulations, with an average backbone root mean squared deviation (bRMSD) of 1.16 Å and 2.36Å for ubiquitin and GB3, respectively. The EEF1-SB trajectories are compared with CHARMM 36 *in vacuo* simulations and with three different implicit solvent models: the EEF1 effective energy function [13], the analytic continuum electrostatics (ACE) model [28], and the GB-based fast analytical continuum treatment of solvation (FACTS) [32]. The results, summarized in table 3, show the RMSD for EEF1-SB to be lower compared to EEF1, ACE and the vacuum simulations, while similar results are obtained with respect to FACTS model. The computational cost associated with this latter model, however, is about 100% higher compared to EEF1-SB.

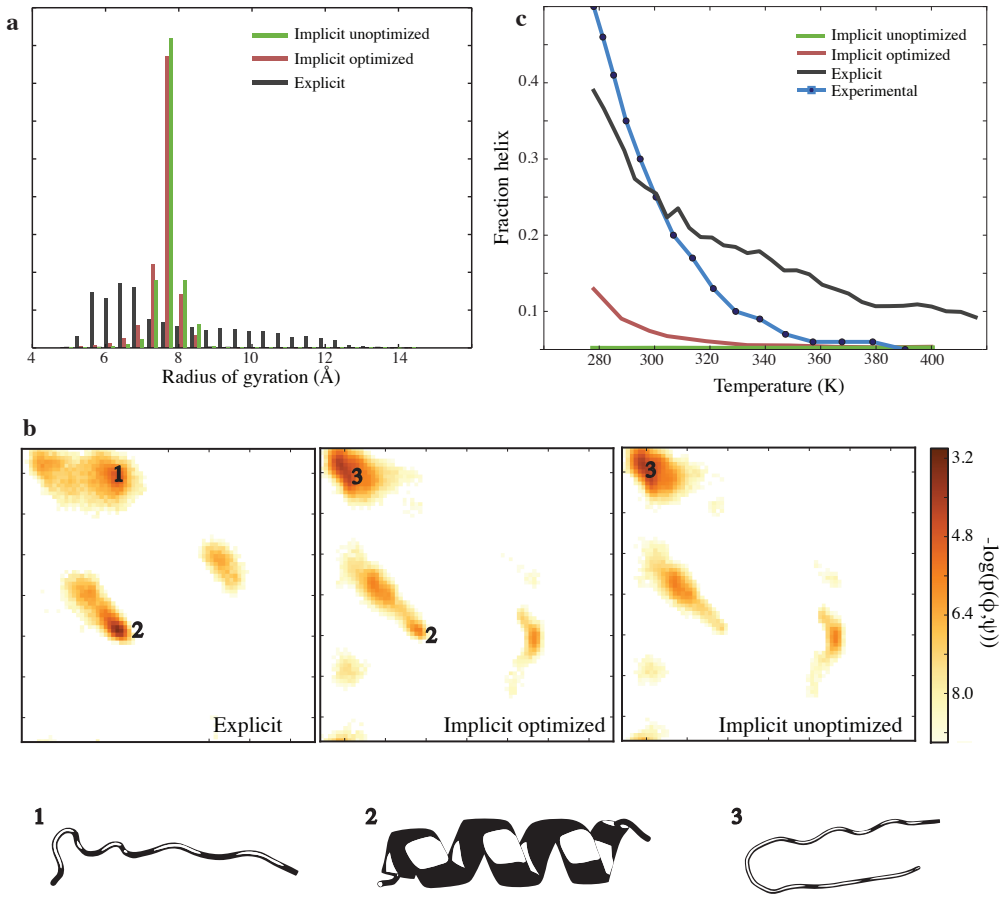


Figure 2. Relative entropy minimization on $(AAQAA)_3$. (**a**) Radius of gyration distribution for the explicit and implicit solvent simulations at T=298K. (**b**) Ramachandran map $(-\log p(\phi, \psi))$ for the 9 central residues and representative structures for three dominant regions. The equilibrium ensemble in explicit water is mainly populated by PPII elongated states (region 1) and α-helical conformations (region 2). The implicit solvent model with optimized parameters favors hairpin-like structures (corresponding to the region labeled with 3 in the Ramachandran map) and α-helical conformations. This latter region is not significantly populated for the unoptimized parameter set. (**c**) Fraction helix for the different solvent models compared to experimental estimate from NMR chemical shifts data [49].

System[a]	Force-field[b]	Solvent model[c]	Rgyr[d] (Å)	bRMSD[e] (Å)	aaRMSD[f]	ns/day [g]
ubq	C19	ACE	10.461	2.304	3.081	2.7
ubq	C19	EEF1	11.129	3.931	5.017	16.7
ubq	C22	FACTS	10.682	0.711	1.427	2.4
ubq	C36	vacuum	10.562	2.200	2.960	6.6
ubq	C36	EEF1-SB	10.939	1.161	2.016	5.6
gb3	C19	ACE	9.998	3.651	4.392	4.3
gb3	C19	EEF1	10.781	4.058	4.787	23.3
gb3	C22	FACTS	10.327	2.159	2.391	4.0
gb3	C36	vacuum	10.419	4.193	4.532	10.5
gb3	C36	EEF1-SB	10.491	2.362	2.625	8.2

Table 3. Simulations on native proteins. (a) pdb codes: ubq (1UBQ) and gb3 (1P7E). (b, c) The different solvent models are used in combination with the associated force-field, as described in the original studies. The united-atom CHARMM 19 force-field (C19) [34] was used for ACE and EEF1, the all-atom force-field CHARMM 22 (C22) [56] for FACTS and the optimized CHARMM 36 (C36) potential [35] for EEF1-SB and vacuum simulations. (d, e, f) Radius of gyration, heavy backbone atoms RMSD (bRMSD) and all-atom (excluding hydrogens) RMSD. For ubiquitin, residues 71-76 are not included in the RMSD calculation.(g) Approximate computational time on a 1.86Ghz Intel Xeon processor expressed in ns/day.

Folding simulations

As a second test, we assess the ability of EEF1-SB to fold different peptides. The aim of this experiment is not to perform a thermodynamic characterization of the system, but rather to assess the free energy minima at room temperature to correspond to conformations compatible to the native structure. Here, three well studied systems with different secondary structure propensities are considered: the β-hairpin of the B1 domain of protein G (which we will refer to as GB1) [57], the α-helical mini protein Trp-cage [58] and the three-stranded β-sheets beta3s peptide [59]. For each system, we report the results from 250ns REMD simulations (16 replicas spanning the temperature range 278-525K) using EEF1-SB. All simulations were initialized from an extended conformation, and the first 100ns were discarded in the analysis.

GB1

The first system we consider is the β-hairpin of GB1 (pdb code 1GB1, residues 41-56). Experimental studies suggest this peptide to populate 40%-60% conformations similar to the β-hairpin of the full structure [57], with a folding time of $6\mu s$ [60]. GB1 has been used in numerous com-

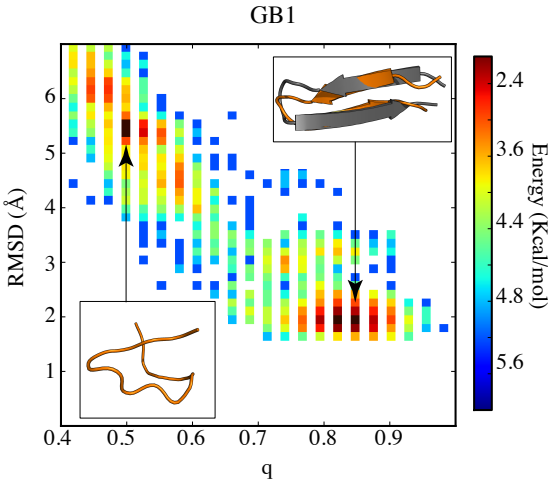


Figure 3. EEF1-SB folding free-energy surface of GB1 at T=298K. The two-dimensional surface is calculated by projecting onto the the number of native contacts q and the backbone RMSD from residues 41-56 of the native structure (pdb code 1GB1). In the insets, two representative structures from the dominant basins are shown in orange. The folded state is superimposed to the experimental structure, shown in dark gray.

putational studies as a model system to investigate the mechanism of hairpin formation [61,62], and has served as a benchmark for force-field validation and comparison [40,41,63,64]. The free energy surface obtained from the EEF1-SB simulation (shown in Fig. 3), reveals two dominant basins: an "unfolded state" at q=0.5 and a "folded" conformation around q=0.9, where q is the number of native CA contacts calculated using a cutoff of 6.5 Å. This result is in line with previous studies where the original EEF1 model was employed on the same system [65], and is in qualitative agreement with explicit water simulation [64].

Trp-cage

Trp-cage is a 20-residues designed protein derived from a fragment of a larger protein. Under physiological conditions, this peptide is $\approx 95\%$ folded, with an estimated folding time of $4\mu s$ [58]. The NMR-derived native structure contains an α-helix, a 3_{10} helix and a polyproline II C-terminus, as shown in Fig. 4. Because of its small size and definite structural propensity, Trp-cage, similarly to GB1, has been the subject of numerous simulation studies, that demonstrated the ability of different force-fields to correctly fold this system with an accuracy up to 1Å

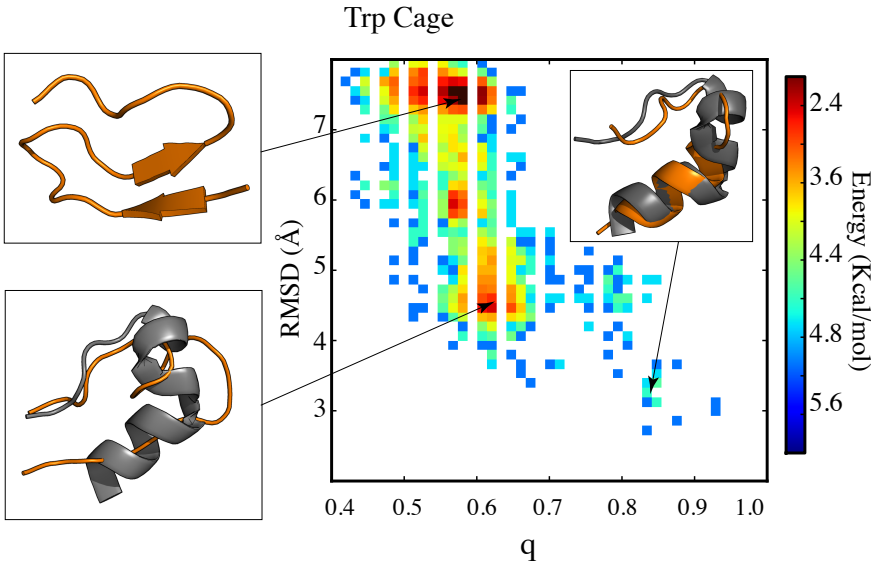


Figure 4. EEF1-SB folding free energy surface of Trp-cage at T=298K. The two-dimensional surface is calculated by projecting onto the the number of native contacts q and the backbone RMSD from native structure (pdb code 1L2Y). In the insets, representative conformations from selected basins are shown in orange, and superimposed to the experimental structure (dark gray). Chain configurations corresponding to the dominant basins are weakly structured, while near-native conformations (RMSD < 3 Å) are not significantly populated.

[41, 63, 66–71]. As shown in Fig. 4, the EEF1-SB model fails to detect the correct fold. The dominant basin is characterized by compact but weakly structured conformations, and near-native states (RMSD < 3 Å) are only weakly populated. In line with the results obtained on the $(AAQAA)_3$ peptide presented in the previous section, the helical content for Trp-cage is negligible, suggesting this missing feature to be responsible for the discrepancy between the simulated and expected free energy surface. Moreover, this sytem has not been tested in explicit solvent simulations and there could be some deficiency also in the underlying CHARMM 36 force-field.

Beta3s

As a last test, we focused on the folding of the 20-residue protein beta3s [59]. Experimental and computational studies suggests beta3s to fold to a defined three-stranded, antiparallel β-

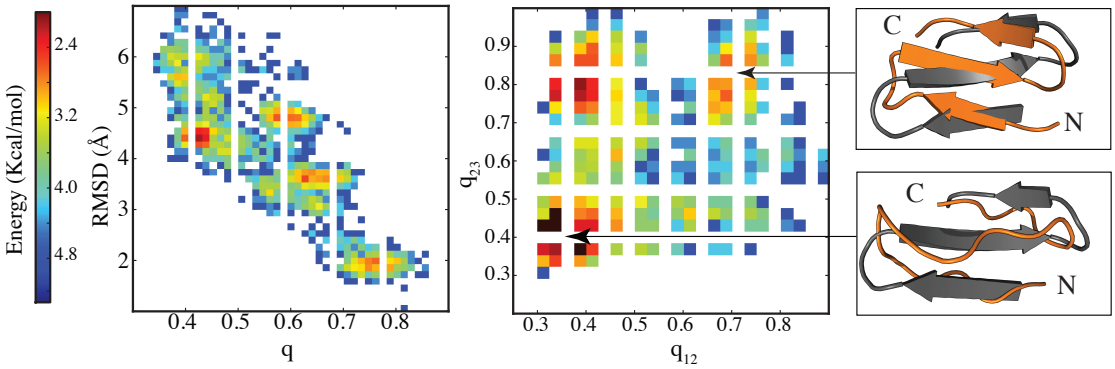


Figure 5. EEF1-SB folding free energy surface of beta3s at T=298K. Left panel: projection onto the the number of native contacts q and the backbone RMSD from the putative native structure. Right panel: free energy surface calculated by projecting onto the the number of formed contacts in the first (q_{12}) and second (q_{23}) hairpin. In the insets, representative structures from two selected basins are shown in orange, and superimposed to the putative native structure.

sheet structure in equilibrium with a heterogeneous ensemble of non-native states [72, 73]. The behavior of beta3s was investigated in a number of different studies by MD simulations in implicit solvent [63,74,75], which showed reversible folding at 330K and a weak temperature dependence of the free energy landscape. Compatibly with previous findings [74], different basins can be identified in the projection of the EEF1-SB trajectory onto the number of contacts in hairpin 1 and hairpin 2 (Fig. 5). More precisely, "folded" states ($q_{12}, q_{23} > 0.7$) are in equilibrium with "unfolded" structures ($q_{12}, q_{23} < 0.4$) and with conformations in which only one of the two hairpins is fully formed.

Discussion

Describing the effect of the aqueous environment is of fundamental importance in molecular simulations of biomolecules. While explicit water simulations provide a high level of detail, implicit solvent models represent a fast and approximate way to describe the behavior of proteins in solution.

The design of an implicit solvent model often entails three distinct but connected tasks. First, an assumption on the physical model describing the effects arising from the presence of the solvent. Secondly, the choice of a force-field describing the solute-solute interactions. It should be noted that most of the recent molecular mechanics force-fields are designed and tested to be used with explicit water models. Therefore, their combination with an implicit solvent representation requires adequate adjustments in the parameters of the resulting effective energy function.

In the present work we employed an approximate, but extremely fast exclusion-solvent model and combined it with the all-atom CHARMM 36 force-field. CHARMM 36 was extensively optimized against experimental data, by performing long molecular dynamics simulations in explicit water. In order to retain the accuracy of the original force-field, we optimized the parameters in the solvent-exclusion model so as to mimic the behavior of explicit water simulations, using a coarse-graining technique based on the minimization of the relative entropy. Although not to a perfect level of detail, the obtained parameters are shown to improve the similarity of the implicit solvent structural ensemble with respect to the explicit solvent equivalent.

Using the optimized parameters, different tests were conducted to ensure that the model gives realistic results. EEF1-SB was shown to give stable native proteins in room temperature molecular dynamics simulations, and reasonable results were obtained from folding studies on β-sheets peptides. Given the poor performance of EEF1-SB on helical systems, a further optimization of the model is desirable. The goal can be achieved using different approaches. One possible route is to make extensive use of reweighting techniques in the relative entropy minimization. This would alleviate some of the computational and numerical problems connected to the procedure, thus allowing a more accurate parameterization. Ideally, only the implicit solvent model would be optimized. However, it may be that certain important properties such as helicity cannot be matched by tuning the solvent model alone. In this case, it may be useful to include additional parameters (e.g. torsion parameters) directly into the relative entropy procedure, but only as a final tuning step once the solvent model has ben optimized

as far as possible. It is also likely that a more precise description of the electrostatic effects is needed to improve the EEF1-SB model. While an accurate treatment of the problem is given by Poisson-Boltzmann and Generalized Born models, these methods are often computationally very expensive, although dramatic speed-ups were reported with the use of graphic processing units (GPU) [76]. Alternative approaches, such as the screened coulomb potential model [77], would instead offer a good compromise between efficiency and accuracy.

Simulation setup

Explicit water simulations The explicit water simulations used in this work are taken from the study of Best *et. al* [35]. For ease of reference, we briefly report the simulation conditions. All simulations were performed with the CHARMM 36 force-field and TIP3P water model using GROMACS 4.5.3 [78]. Long range electrostatics were treated using Particle-Mesh Ewald [79] summation with a real-space cutoff of 12 Å and a 1 Å grid spacing, while the Lennard-Jones interactions were treated with a switching function from 10 to 12 Å. The equations of motion were integrated with a 2 fs time step. Bond lengths were constrained using the SHAKE algorithm [80], while SETTLE [81] maintained rigid water geometries. The 100ns simulation on the unblocked Ala5 peptide with protonated C-terminal was performed in the NPT ensemble at 298K and 1 atm pressure in a box of size 34.56 $Å^3$. A Langevin thermostat and barostat were used. The Ac-$(AAQAA)_3$-NH_2 peptide was solvated with 1833 water molecules in a truncated octahedron cell with a distance between nearest faces of 42 Å. The peptide was first unfolded using a 5 ns constant volume simulation at 800 K. Subsequently, a constant volume replica exchange MD was run, with 32 replicas spanning a temperature range from 278 to 416 K and exchange attempts every 10 ps, for a total of 150 ns per replica, of which only the last 50ns were used for the relative entropy procedure. A Langevin thermostat with a friction coefficient of 1 ps^{-1} was used.

Implicit water simulations Simulations in implicit water were performed with the CHARMM 36 force-field using the CHARMM software package [82]. Langevin dynamics with a friction coefficient of 1 ps^{-1} was used, and all bond lengths were constrained using the SHAKE algorithm. REMD simulations were performed spanning different temperature ranges, as described in the main text, and with exchange attempts every 0.4ps.

References

1. Dill K (1990) Dominant forces in protein folding. Biochemistry 29: 7133–7155.

2. Prabhu N, Sharp K (2006) Protein-solvent interactions. Chemical Reviews 106: 1616–1623.

3. Bryant R (1996) The dynamics of water-protein interactions. Annual Review of Biophysics and Biomolecular Structure 25: 29–53.

4. Tarek M, Tobias D (2000) The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. Biophysical Journal 79: 3244–3257.

5. Zhang L, Wang L, Kao Y, Qiu W, Yang Y, et al. (2007) Mapping hydration dynamics around a protein surface. Proceedings of the National Academy of Sciences 104: 18461–18466.

6. Eisenberg D, McLachlan A, et al. (1986) Solvation energy in protein folding and binding. Nature 319: 199–203.

7. Ooi T, Oobatake M, Nemethy G, Scheraga H (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proceedings of the National Academy of Sciences 84: 3086–3090.

8. Fraternali F, Van Gunsteren W (1996) An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. Journal of Molecular Biology 256: 939–948.

9. Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins: Structure, Function, and Bioinformatics 46: 24–33.

10. Colonna-Cesari F, Sander C (1990) Excluded volume approximation to protein-solvent interaction. the solvent contact model. Biophysical Journal 57: 1103–1107.

11. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophysical Journal 88: 1560–1569.

12. Stouten P, Frömmel C, Nakamura H, Sander C (1993) An effective solvation term based on atomic occupancies for use in protein simulations. Molecular Simulation 10: 97–120.

13. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins: Structure, Function, and Bioinformatics 35: 133–152.

14. Stultz C (2004) An assessment of potential of mean force calculations with implicit solvent models. The Journal of Physical Chemistry B 108: 16525–16532.

15. Sun Y, Latour R (2006) Comparison of implicit solvent models for the simulation of protein–surface interactions. Journal of Computational Chemistry 27: 1908–1922.

16. Pu M, Garrahan J, Hirst J (2011) Comparison of implicit solvent models and force fields in molecular dynamics simulations of the pb1 domain. Chemical Physics Letters 515: 283–289.

17. Inuzuka Y, Lazaridis T (2000) On the unfolding of α-lytic protease and the role of the pro region. Proteins: Structure, Function, and Bioinformatics 41: 21–32.

18. Hassan S, Mehler E (2002) A critical analysis of continuum electrostatics: the screened coulomb potential–implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. Proteins: Structure, Function, and Bioinformatics 47: 45–61.

19. Huang A, Stultz C (2007) Conformational sampling with implicit solvent models: application to the phf6 peptide in tau protein. Biophysical Journal 92: 34–45.

20. Cavalli A, Salvatella X, Dobson C, Vendruscolo M (2007) Protein structure determination from nmr chemical shifts. Proceedings of the National Academy of Sciences 104: 9615-9621.

21. Kaufmann K, Lemmon G, DeLuca S, Sheehan J, Meiler J (2010) Practically useful: what the rosetta protein modeling suite can do for you. Biochemistry 49: 2987–2998.

22. Ding F, Tsao D, Nie H, Dokholyan N (2008) Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure 16: 1010–1018.

23. Warwicker J, Watson H (1982) Calculation of the electric potential in the active site cleft due to α-helix dipoles. Journal of Molecular Biology 157: 671–679.

24. Nicholls A, Honig B (1991) A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson–boltzmann equation. Journal of Computational Chemistry 12: 435–445.

25. Luo R, David L, Gilson M (2002) Accelerated poisson–boltzmann calculations for static and dynamic systems. Journal of Computational Chemistry 23: 1244–1253.

26. Constanciel R (1986) Theoretical basis of the empirical reaction field approximations through continuum model. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 69: 505–523.

27. Still W, Tempczyk A, Hawley R, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. Journal of the American Chemical Society 112: 6127–6129.

28. Schaefer M, Karplus M (1996) A comprehensive analytical treatment of continuum electrostatics. The Journal of Physical Chemistry 100: 1578–1599.

29. Di Qiu M, Shenkin P, Hollinger F, Still W (1997) The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. The Journal of Physical Chemistry A 101: 3005–3014.

30. Lee M, Salsbury F, Brooks C (2002) Novel generalized born methods. Journal of Chemical Physics 116: 10606–10614.

31. Im W, Lee M, Brooks III C (2003) Generalized born model with a simple smoothing function. Journal of Computational Chemistry 24: 1691–1702.

32. Haberthür U, Caflisch A (2008) Facts: fast analytical continuum treatment of solvation. Journal of Computational Chemistry 29: 701–715.

33. Feig M, Reiher M, Wolf A (2010) Modeling solvent environments. Wiley Online Library.

34. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. The Journal of Chemical Physics 105: 1902–1922.

35. Best RB, Zhu X, Shim J, Lopes P, Mittal J, et al. (2012) Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ,ψ and side-chain χ_1 and χ_2 dihedral angles. Unpublished manuscript .

36. Shell M (2008) The relative entropy is fundamental to multiscale and inverse thermodynamic problems. The Journal of Chemical Physics 129: 144108–144115.

37. Chaimovich A, Shell M (2010) Relative entropy as a universal metric for multiscale errors. Physical Review E 81: 060104–060108.

38. Chaimovich A, Shell M (2011) Coarse-graining errors and numerical optimization using a relative entropy framework. The Journal of Chemical Physics 134: 094112–0941127.

39. Carmichael S, Shell M (2012) A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. The Journal of Physical Chemistry B .

40. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood M, Dror R, et al. (2012) Systematic validation of protein force fields against experimental data. PloS one 7: e32131.

41. Best R, Mittal J (2010) Balance between α and β structures in ab initio protein folding. The Journal of Physical Chemistry B 114: 8790–8798.

42. Privalov P, Makhatadze G (1993) Contribution of hydration to protein folding thermodynamics. the entropy and gibbs energy of hydration. Journal of Molecular Biology 232: 660–679.

43. Makhatadze G, Privalov P (1993) Contribution of hydration to protein folding thermodynamics:: I. the enthalpy of hydration. Journal of Molecular Biology 232: 639–659.

44. Privalov P, Makhatadze G (1992) Contribution of hydration and non-covalent interactions to the heat capacity effect on protein unfolding. Journal of Molecular Biology 224: 715–723.

45. Jorgensen W (1981) Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. Journal of the American Chemical Society 103: 335–340.

46. Zwanzig R (1954) High-temperature equation of state by a perturbation method. i. nonpolar gases. The Journal of Chemical Physics 22: 1420–1427.

47. Norgaard A, Ferkinghoff-Borg J, Lindorff-Larsen K (2008) Experimental parameterization of an energy function for the simulation of unfolded proteins. Biophysical journal 94: 182–192.

48. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314: 141–151.

49. Shalongo W, Dugad L, Stellwagen E (1994) Distribution of helicity within the model peptide acetyl (aaqaa) 3amide. Journal of the American Chemical Society 116: 8288–8293.

50. Best R, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. The Journal of Physical Chemistry B 113: 9004–9015.

51. Piana S, Lindorff-Larsen K, Shaw D, et al. (2011) How robust are protein folding simulations with respect to force field parameterization? Biophysical Journal 100: 47.

52. Rucker A, Creamer T (2002) Polyproline ii helical structure in protein unfolded states: Lysine peptides revisited. Protein Science 11: 980–985.

53. Liu Z, Chen K, Ng A, Shi Z, Robert W, et al. (2004) Solvent dependence of pii conformation in model alanine peptides. Journal of the American Chemical Society 126: 15141–15150.

54. Markwick P, Bouvignies G, Blackledge M (2007) Exploring multiple timescale motions in protein gb3 using accelerated molecular dynamics and nmr spectroscopy. Journal of the American Chemical Society 129: 4724–4730.

55. Lange O, Lakomek N, Farès C, Schröder G, Walter K, et al. (2008) Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. Science 320: 1471–1475.

56. MacKerell Jr A, Bashford D, Bellott M, Dunbrack Jr R, Evanseck J, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. The Journal of Physical Chemistry B 102: 3586–3616.

57. Blanco F, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nature Structural & Molecular Biology 1: 584–590.

58. Qiu L, Pabit S, Roitberg A, Hagen S (2002) Smaller and faster: the 20-residue trp-cage protein folds in 4 μs. Journal of the American Chemical Society 124: 12952–12953.

59. De Alba E, Santoro J, Rico M, Jimenez M (1999) De novo design of a monomeric three-stranded antiparallel [beta]-sheet. Protein Science 8: 854–865.

60. Munoz V, Thompson P, Hofrichter J, Eaton W (1997) Folding dynamics and mechanism of b-hairpin formation. Nature 390: 196–198.

61. Zagrovic B, Sorin E, Pande V (2001) [beta]-hairpin folding simulations in atomistic detail using an implicit solvent model. Journal of Molecular Biology 313: 151–169.

62. Nguyen P, Stock G, Mittag E, Hu C, Li M (2005) Free energy landscape and folding mechanism of a β-hairpin in explicit water: A replica exchange molecular dynamics study. PROTEINS: Structure, Function, and Bioinformatics 61: 795–808.

63. Irbäck A, Mitternacht S, Mohanty S (2009) An effective all-atom potential for proteins. BMC Biophysics 2: 2.

64. Best R, Mittal J (2011) Free-energy landscape of the gb1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences. Proteins: Structure, Function, and Bioinformatics 79: 1318–1328.

65. Dinner A, Lazaridis T, Karplus M (1999) Understanding β-hairpin formation. Proceedings of the National Academy of Sciences 96: 9068–9074.

66. Snow C, Zagrovic B, Pande V (2002) The trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. Journal of the American Chemical Society 124: 14548–14549.

67. Zhou R (2003) Trp-cage: folding free energy landscape in explicit water. Proceedings of the National Academy of Sciences of the United States of America 100: 13280–13286.

68. Chowdhury S, Lee M, Xiong G, Duan Y (2003) Ab initio folding simulation of the trp-cage mini-protein approaches nmr resolution. Journal of molecular biology 327: 711–717.

69. Juraszek J, Bolhuis P (2006) Sampling the multiple folding mechanisms of trp-cage in explicit solvent. Proceedings of the National Academy of Sciences 103: 15859–15864.

70. Chen J, Im W, Brooks III C (2006) Balancing solvation and intramolecular interactions: toward a consistent generalized born force field. Journal of the American Chemical Society 128: 3728–3736.

71. Paschek D, Nymeyer H, García A (2007) Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. Journal of Structural Biology 157: 524–533.

72. Cavalli A, Haberthür U, Paci E, Caflisch A (2003) Fast protein folding on downhill energy landscape. Protein Science 12: 1801–1803.

73. Krivov S, Muff S, Caflisch A, Karplus M (2008) One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. The Journal of Physical Chemistry B 112: 8701–8714.

74. Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. Proceedings of the National Academy of Sciences 97: 10780.

75. Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. Proteins: Structure, Function, and Bioinformatics 47: 305–314.

76. Friedrichs M, Eastman P, Vaidyanathan V, Houston M, Legrand S, et al. (2009) Accelerating molecular dynamic simulation on graphics processing units. Journal of Computational Chemistry 30: 864–872.

77. Hassan S, Guarnieri F, Mehler E (2000) A general treatment of solvent effects based on screened coulomb potentials. The Journal of Physical Chemistry B 104: 6478–6489.

78. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of Chemical Theory and Computation 4: 435–447.

79. York D, Darden T, Pedersen L (1993) The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the ewald and truncated list methods. The Journal of Chemical Physics 99: 8345–8349.

80. Ryckaert J, Ciccotti G, Berendsen H (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics 23: 327–341.

81. Miyamoto S, Kollman P (1992) Settle: an analytical version of the shake and rattle algorithm for rigid water models. Journal of Computational Chemistry 13: 952–962.

82. Brooks B, Brooks III C, Mackerell Jr A, Nilsson L, Petrella R, et al. (2009) Charmm: the biomolecular simulation program. Journal of Computational Chemistry 30: 1545–1614.

3

Conclusion

3.1 Concluding Remarks

One of the main challenges connected to computer simulations of biomolecules is to extend the *observation time* (*i.e.* the length of the simulation) to the typical time-scale over which the biological process of interest occurs. This problem is very relevant: calculating average quantities over unconverged simulations can lead to dramatic errors, which would therefore give wrong insights into the problem under consideration. It should also be noted that accessing long time-scales allows for a direct comparison between results from simulations and experimental measurements, thus providing a robust route for validating the employed computational approach. It therefore comes as no surprise that a substantial effort of the scientific community is devoted to the development of high-performance hardware, efficient software and smart sampling algorithms.

In this dissertation I presented a number of methodologies for efficient conformational sampling of proteins. Specifically, I introduced a novel Monte Carlo algorithm, CRISP, and derived a coarse-grained representation for modeling protein-solvent interactions. When possible, I compared the performance of these methods to the current state-of-the-art techniques and with experimental data. As the simulation community grows, it becomes more and more important to perform a thorough assessment of new methodologies, in order to make the results of the study useful to others in the field.

The main limitation of the approaches presented in this dissertation is connected to the use of implicit solvent models. The combination of the $OPLS_{AA}$ force-field with

the generalized Born/surface area (GB/SA) model used in the study of section 2.1 has been criticized for over-stabilizing salt bridges [146]. Moreover, the computational cost of GB approaches on standard CPUs is only marginally lower compared to explicit water simulations, although a dramatic speed-up can be achieved by using graphic processing units (GPU) [147]. On the other hand, the effective EEF1/EEF1-SB models are extremely fast, but not always accurate. A good compromise between accuracy and efficiency can be in principle obtained by combining the EEF1-SB implicit solvent with a simple model that accounts for electrostatic effects, such as the screened Coulomb potential approach [148]. Bearing in mind the aforementioned disadvantages, implicit solvent can be used to study a wide range of biological systems, and, as a large body of work demonstrates [25]-[29], it can be fruitfully used in combination with Monte Carlo simulations.

Other aspects of the work presented in this dissertation can be improved. Specifically, future developments of the software package PHAISTOS (section 2.2) could progress roughly along three different lines.

First, the complementarity with molecular dynamics methods. From a theoretical standpoint it is important to show that MC and molecular dynamics give comparable results, even the in dense environment of native globular proteins (see section 2.1). Application-wise, it is however more natural to use MC for studying systems that undergo large structural or spatial rearrangements, such as intrinsically disordered proteins or aggregation-prone peptides. Computational studies of these processes require specific force-field and the possibility of handling multiple chains, but both features are currently not available within the package. The use of experimental data to guide the MC simulation is another promising approach that we partially considered (section 2.4 and Ref. [149]) and that recently became very popular within the field [150].

The ease of use is a second aspect that could be improved. As also discussed in section 1.2, the efficiency of an MC simulation relies on the choice of the move and on the associated parameters. In turn, the optimal settings depend on the system of interest, on the specific problem and on the desired level of detail. Clearly, this leaves the user of the software with a large set of free parameters. One possible solution to this problem is the introduction of an automated system for optimizing the move type/size, along the same lines as the approach used in the CHARMM MC module [151].

Finally, further extensions and refinements to the move-set can be introduced. In the Monte Carlo move presented in section 2.1, side-chain and backbone degrees of freedom are treated independently. This approach, although very common, neglects completely the fact that even a small, local backbone variation can lead to a large displacement of the rigidly attached side-chain, and therefore to the rejection of the trial configuration. Similarly, uncorrelated motions of interacting side-chains can be energetically unfavorable (*e.g.* breaking of hydrogen bonds). In this context, the Gaussian-biased step approach [37] and CRISP moves can be combined in a large number of different ways to further enhance the efficiency of MC simulations (Figure 3.1).

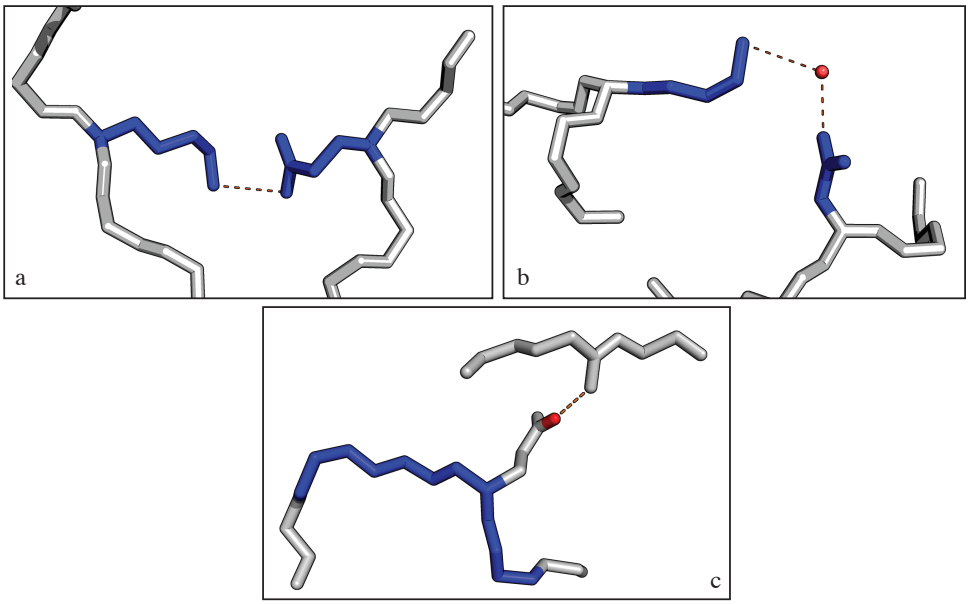


Figure 3.1: Extended CRISP - (**a**-**b**) CRISP moves can be applied for side-chain sampling, by constructing a fictitious chain connecting interacting side-chains or even involving one or more water molecules. This type of move allows to vary the internal degrees of freedom, while preserving the polar interactions (dashed line). The region of the chain affected by the local update is colored in blue. (**c**) Alternatively, it is possible to introduce purely backbone moves with a bias toward a small displacement of a side-chain endpoint (colored in red).

3.2 Acknowledgements

I would like to express my gratitude to all the people who, directly and indrectly, helped and supported me during these three long, exciting years. First of all I would like to thank my supervisor Jesper, who introduced me into the joyful world of Science, and guided me from the very first day with his fantastic, everlasting enthusiasm. A special thank goes to Wouter Boomsma: thanks your invaluable help and countless discussions (actually not really countless: my inbox says 1065 conversations over the last 3 years). I have enjoyed very much working with you.

The whole group of structural bioinformatics at the University of Copenhagen: Christian Andreetta, Jes Frellsen, Kristoffer Enøe Johansson, Tim Harder (you're such a good sport too!), Simon Olsson, Jan Valentin, Mikael Borg and Kasper Stovgaard. Thanks for the good old table football sessions and for introducing me to the importance in the Danish culture of the fiskefilet-based early lunch at 11.30. A special thank goes to Thomas Hamelryck for the many good suggestions and for always making me feel welcome in your group. I would also like to express my gratitude the people in Cambridge for the great hospitality. In particular, to Robert Best and David de Sancho for the time spent helping me. Without Kresten Lindorff-Larsen much of my reaserch during the last year would not have been possible: thanks for the guidance and for your valuable support.

My friends, in particular Claudio, Luigi, Giuseppe, Alessandro and Jacopo for always reminding me that there is a life outside the department. I am deeply thankful to my mother (mamma!), to my father, to my brother Angelo, toghether with Elena and the little Sofia for always having been supportive of what I am doing. Last, but by no means least, thanks to my fantastic girlfriend Claudia, for your love and your constant presence despite the long distance.

Det var hyggeligt!

Sandro Bottaro,
Copenhagen, April 2012

References

[1] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. **Equation of state calculations by fast computing machines**. *The Journal of Chemical Physics*, **21**(6):1087, 1953. 1, 6

[2] B.J. Alder and TE Wainwright. **Studies in molecular dynamics. I. General method**. *The Journal of Chemical Physics*, **31**(2):459, 1959. 1

[3] J.L. Klepeis, K. Lindorff-Larsen, R.O. Dror, and D.E. Shaw. **Long-timescale molecular dynamics simulations of protein structure and function**. *Current Opinion in Structural Biology*, **19**(2):120, 2009. 1

[4] K. Lindorff-Larsen, S. Piana, R.O. Dror, and D.E. Shaw. **How fast-folding proteins fold**. *Science's STKE*, **334**(6055):517, 2011. 1

[5] K. Henzler-Wildman and D. Kern. **Dynamic personalities of proteins**. *Nature*, **450**(7172):964, 2007. 2

[6] K. Teilum, J.G. Olsen, and B.B. Kragelund. **Functional aspects of protein flexibility**. *Cellular and Molecular Life Sciences*, **66**(14):2231, 2009. 3

[7] G.A. Jensen, O.M. Andersen, A.M.J.J. Bonvin, I. Bjerrum-Bohr, M. Etzerodt, H.C. Thøgersen, C. O'Shea, F.M. Poulsen, and B.B. Kragelund. **Binding site structure of one LRP–RAP complex: Implications for a common ligand–receptor binding motif**. *Journal of Molecular Biology*, **362**(4):700, 2006. 4

[8] J. Kubelka, J. Hofrichter, and W.A. Eaton. **The protein folding speed limit'**. *Current Opinion in Structural Biology*, **14**(1):76, 2004. 4

[9] J.A. McCammon, B.R. Gelin, M. Karplus, et al. **Dynamics of folded proteins**. *Nature*, **267**(5612):585, 1977. 4

[10] A. Li and V. Daggett. **Investigation of the solution structure of chymotrypsin inhibitor 2 using molecular dynamics: comparison to X-ray crystallographic and NMR data**. *Protein Engineering*, **8**(11):1117, 1995. 4

[11] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, et al. **Atomic-level characterization of the structural dynamics of proteins**. *Science*, **330**(6002):341, 2010. 4

[12] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier (formely published by Academic Press), 1996. 3

[13] W.L. Jorgensen, D.S. Maxwell, and J. Tirado-Rives. **Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids**. *Journal of the American Chemical Society*, **118**(45):11225, 1996. 5

[14] A. Irbäck and S. Mohanty. **PROFASI: a Monte Carlo simulation package for protein folding and aggregation**. *Journal of Computational Chemistry*, **27**(13):1548, 2006. 5

[15] A. Vitalis and R.V. Pappu. **Methods for Monte Carlo simulations of biomacromolecules**. *Annual Reports in Computational Chemistry*, **5**(6):49, 2009. 5, 16

[16] W.L. Jorgensen and J. Tirado-Rives. **Molecular modeling of organic and biomolecular systems using BOSS and MCPRO**. *Journal of Computational Chemistry*, **26**(16):1689, 2005. 5

[17] E. Lindahl, B. Hess, and D. Van Der Spoel. **GROMACS 3.0: a package for molecular simulation and trajectory analysis**. *Journal of Molecular Modeling*, **7**(8):306, 2001. 5

[18] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. **Scalable molecular dynamics with NAMD**. *Journal of Computational Chemistry*, **26**(16):1781, 2005. 5

[19] D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. **The Amber biomolecular simulation programs**. *Journal of Computational Chemistry*, **26**(16):1668, 2005. 5

[20] B.R. Brooks, CL Brooks III, AD Mackerell Jr, L. Nilsson, RJ Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. **CHARMM: the biomolecular simulation program**. *Journal of Computational Chemistry*, **30**(10):1545, 2009. 5

[21] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. **Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes**. *Journal of Computational Physics*, **23**(3):327, 1977. 5

[22] H.C. Andersen. **Molecular dynamics simulations at constant pressure and/or temperature**. *The Journal of Chemical Physics*, **72**(4):2384, 1980. 5

[23] C. Bartels and M. Karplus. **Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations**. *Journal of Computational Chemistry*, **18**(12):1450, 1997. 5

[24] Y. Sugita and Y. Okamoto. **Replica-exchange molecular dynamics method for protein folding**. *Chemical Physics Letters*, **314**(1):141, 1999. 5

[25] S. Mitternacht, I. Staneva, T. Hard, and A. Irback. **Monte Carlo study of the formation and conformational properties of dimers of A [beta] 42 variants**. *Journal of Molecular Biology*, **410**(2), 2011. 5, 124

[26] D.W. Li, S. Mohanty, A. Irbäck, and S. Huo. **Formation and growth of oligomers: a Monte Carlo study of an amyloid tau fragment**. *PLoS Computational Biology*, **4**(12):e1000238, 2008. 5

[27] A. Vitalis, X. Wang, and R.V. Pappu. **Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization**. *Journal of Molecular Biology*, **384**(1):279, 2008. 5

[28] J.P. Ulmschneider, M.B. Ulmschneider, and A. Di Nola. **Monte carlo folding of trans-membrane helical peptides in an implicit generalized Born membrane**. *Proteins: Structure, Function, and Bioinformatics*, **69**(2):297, 2007. 5

[29] J.P. Ulmschneider and M.B. Ulmschneider. **Folding simulations of the transmembrane helix of virus protein U in an implicit membrane model**. *Journal of Chemical Theory and Computation*, **3**(6):2335, 2007. 5, 124

REFERENCES

[30] J.P. Ulmschneider, M.B. Ulmschneider, and A. Di Nola. **Monte Carlo vs molecular dynamics for all-atom polypeptide folding simulations**. *The Journal of Physical Chemistry B*, **110**(33):16733, 2006. 5

[31] S.H. Northrup and J.A. McCammon. **Simulation methods for protein structure fluctuations**. *Biopolymers*, **19**(5):1001, 1980. 5

[32] T. Noguti and N. Gō. **Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins**. *Biopolymers*, **24**(3):527, 1985. 5, 16

[33] A. Kidera and N. Go. **Refinement of protein dynamic structure: normal mode refinement**. *Proceedings of the National Academy of Sciences*, **87**(10):3718, 1990. 5

[34] A. Irbäck, S. Mitternacht, and S. Mohanty. **Dissecting the mechanical unfolding of ubiquitin**. *Proceedings of the National Academy of Sciences*, **102**(38):13427, 2005. 5

[35] M Lal. **Monte Carlo computer simulations of chain molecules**. *Molecular Physics*, **17**(57):64, 1969. 6

[36] L. Holm and C. Sander. **Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology**. *Proteins: Structure, Function, and Bioinformatics*, **14**(2):213, 1992. 6

[37] G. Favrin, A. Irbäck, and F. Sjunnesson. **Monte Carlo update for chain molecules: biased Gaussian steps in torsional space**. *The Journal of Chemical Physics*, **114**(18):8154, 2001. 7, 11, 12, 125

[38] P.G. de Gennes. **Reptation of a polymer chain in the presence of fixed obstacles**. *The Journal of Chemical Physics*, **55**(2):572, 1971. 9

[39] A.A. Jones and W.H. Stockmayer. **Models for spin relaxation in dilute solutions of randomly coiled polymers**. *Journal of Polymer Science: Polymer Physics Edition*, **15**(5):847, 1977. 9

[40] S.K. Kumar, M. Vacatello, and D.Y. Yoon. **Off-lattice Monte Carlo simulations of polymer melts confined between two plates**. *The Journal of Chemical Physics*, **89**(8):5206, 1988. 9

[41] M.R. Betancourt. **Efficient Monte Carlo trial moves for polypeptide simulations**. *The Journal of Chemical Physics*, **123**(17):174905, 2005. 9

[42] C.A. Smith and T. Kortemme. **Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction**. *Journal of Molecular Biology*, **380**(4):742, 2008. 9

[43] R.A. Engh and R. Huber. **Accurate bond and angle parameters for X-ray protein structure refinement**. *Acta Crystallographica Section A: Foundations of Crystallography*, **47**(4):392, 1991. 9

[44] N. Gō and H.A. Scheraga. **Ring closure and local conformational deformations of chain molecules**. *Macromolecules*, **3**(2):178, 1970. 9, 10

[45] LR Dodd, TD Boone, and DN Theodorou. **A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses**. *Molecular Physics*, **78**(4):961, 1993. 10

[46] W.J. Wedemeyer and H.A. Scheraga. **Exact analytical loop closure in proteins using polynomial equations**. *Journal of Computational Chemistry*, **20**(8):819, 1999. 10

[47] D. Manocha, Y. Zhu, and W. Wright. **Conformational analysis of molecular chains using nano-kinematics**. *Computer Applications in the Biosciences: CABIOS*, **11**(1):71, 1995. 10

[48] E.A. Coutsias, C. Seok, M.P. Jacobson, and K.A. Dill. **A kinematic view of loop closure**. *Journal of Computational Chemistry*, **25**(4):510–528, 2004. 10

[49] EW Knapp and A. Irgens-Defregger. **Off-lattice Monte Carlo method with constraints: Long-time dynamics of a protein model without nonbonded interactions**. *Journal of Computational Chemistry*, **14**(1):19, 1993. 11

[50] P.V.K. Pant and D.N. Theodorou. **Variable connectivity method for the atomistic Monte Carlo simulation of polydisperse polymer melts**. *Macromolecules*, **28**(21):7224, 1995. 11

[51] V.G. Mavrantzas, T.D. Boone, E. Zervopoulou, and D.N. Theodorou. **End-bridging Monte Carlo: A fast algorithm for atomistic simulation of condensed phases of long polymer chains**. *Macromolecules*, **32**(15):5072, 1999. 11

[52] S. Santos, UW Suter, M. Müller, and J. Nievergelt. **A novel parallel-rotation algorithm for atomistic Monte Carlo simulation of dense polymer systems**. *The Journal of Chemical Physics*, **114**(22):9772, 2001. 11

[53] M Mezei. **Efficient Monte Carlo sampling for long molecular chains using local moves, tested on a solvated lipid bilayer**. *Journal of Chemical Physics*, **118**(8):3874, 2003. 11

[54] D. Hoffmann and E.W. Knapp. **Protein dynamics with off-lattice Monte Carlo moves**. *Physical Review E*, **53**(4):4221, 1996. 11

[55] D. Hoffmann and E.W. Knapp. **Polypeptide folding with off-lattice Monte Carlo dynamics: the method**. *European Biophysics Journal*, **24**(6):387, 1996. 11

[56] A.R. Dinner. **Local deformations of polymers with nonplanar rigid main-chain internal coordinates**. *Journal of Computational Chemistry*, **21**(13):1132, 2000. 11

[57] H. Sklenar, D. Wüstner, and R. Rohs. **Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians**. *Journal of Computational Chemistry*, **27**(3):309, 2006. 11

[58] J.P. Ulmschneider and W.L. Jorgensen. **Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias**. *The Journal of Chemical Physics*, **118**(9):4261, 2003. 11, 13

[59] M.N. Rosenbluth and A.W. Rosenbluth. **Monte Carlo calculation of the average extension of molecular chains**. *The Journal of Chemical Physics*, **23**(2):356, 1955. 11

[60] J.I. Siepmann and D. Frenkel. **Configurational bias Monte Carlo: a new sampling scheme for flexible chains**. *Molecular Physics*, **75**(1):59, 1992. 11

[61] D. Frenkel, G. Mooij, and B. Smit. **Novel scheme to study structural and thermal properties of continuously deformable molecules**. *Journal of Physics: Condensed Matter*, **4**(12):3053, 1992. 11

[62] J.J. De Pablo, M. Laso, and U.W. Suter. **Simulation of polyethylene above and below the melting point**. *The Journal of Chemical Physics*, **96**(3):2395, 1992. 11

[63] F. A. Escobedo and J. J. de Pablo. **Extended continuum configurational bias Monte Carlo methods for simulation of flexible molecules**. *Journal of Chemical Physics*, **102**(6):2637, 1994. 11

[64] M.W. Deem and J.S. Bader. **A configurational bias Monte Carlo method for linear and cyclic peptides**. *Molecular Physics*, **87**(6):1245, 1996. 13

[65] M.G. Wu and Deem. M.W. **Efficient Monte Carlo methods for cyclic peptides**. *Molecular Physics*, **97**(4):559, 1999. 13

[66] Z. Chen and F.A Escobedo. **A configurational-bias approach for the simulation of inner sections of linear and cyclic molecules**. *The Journal of Chemical Physics*, **113**(24):11382, 2000. 13

[67] A Uhlherr. **Monte Carlo conformational sampling of the internal degrees of freedom of chain molecules**. *Macromolecules*, **33**(4):1351, 2000. 13

[68] A. Uhlherr, V.G. Mavrantzas, M Doxastakis, and D.N. Theodorou. **Directed bridging methods for fast atomistic Monte Carlo simulations of bulk polymers**. *Macromolecules*, **34**(24):8554, 2001. 13

[69] M. Vendruscolo. **Modified configurational bias monte carlo method for simulation of polymer systems**. *The Journal of Chemical Physics*, **106**(7):2970, 1997. 13

[70] C.D. Wick and J.I. Siepmann. **Self-adapting fixed-endpoint configurational-bias Monte Carlo method for the regrowth of interior segments of chain molecules with strong intramolecular interactions**. *Macromolecules*, **33**(19):7207, 2000. 13

[71] G.N. Ramachandran and V. Sasisekharan. **Conformation of polypeptides and proteins**. *Advances in Protein Chemistry*, **23**(283):438, 1968. 14

[72] E.G. Hutchinson and J.M. Thornton. **PROMOTIF–a program to identify and analyze structural motifs in proteins.** *Protein Science: A Publication of the Protein Society*, **5**(2):212, 1996. 14

[73] J. Janin, S. WodakMichael, and B. Maigret. **Conformation of amino acid side-chains in proteins**. *Journal of Molecular Biology*, **125**(3):357, 1978. 14

[74] R. Unger, D. Harel, S. Wherland, and J.L. Sussman. **A 3D building blocks approach to analyzing and predicting structure of proteins**. *Proteins: Structure, Function, and Bioinformatics*, **5**(4):355, 1989. 14

[75] C. Bystroff and D. Baker. **Prediction of local structure in proteins using a library of sequence-structure motifs1**. *Journal of Molecular Biology*, **281**(3):565, 1998. 14

[76] R.L. Dunbrack, M. Karplus, et al. **Backbone-dependent rotamer library for proteins application to side-chain prediction**. *Journal of Molecular Biology*, **230**(2):543, 1993. 14

[77] S.C. Lovell, J.M. Word, J.S. Richardson, and D.C. Richardson. **The penultimate rotamer library**. *Proteins: Structure, Function, and Bioinformatics*, **40**(3):389, 2000. 14

[78] J.U. Bowie and D. Eisenberg. **An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function**. *Proceedings of the National Academy of Sciences*, **91**(10):4436, 1994. 14

[79] K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions1**. *Journal of Molecular Biology*, **268**(1):209, 1997. 14

[80] M.J. Bower, F.E. Cohen, R.L. Dunbrack, et al. **Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool1**. *Journal of Molecular Biology*, **267**(5):1268, 1997. 14

[81] AC Camproux, P. Tuffery, JP Chevrolat, JF Boisvieux, and S. Hazout. **Hidden Markov model approach for identifying the modular framework of the protein backbone**. *Protein Engineering*, **12**(12):1063, 1999. 14

[82] C. Bystroff, V. Thorsson, and D. Baker. **HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins1**. *Journal of Molecular Biology*, **301**(1):173, 2000. 14

[83] T. Hamelryck, J.T. Kent, and A. Krogh. **Sampling realistic protein conformations using local structural bias**. *PLoS Computational Biology*, **2**(9):e131, 2006. 14

[84] W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. **A generative, probabilistic model of local protein structure**. *Proceedings of the National Academy of Sciences*, **105**(26):8932, 2008. 14, 15

[85] J. Frellsen, I. Moltke, M. Thiim, K.V. Mardia, J. Ferkinghoff-Borg, and T. Hamelryck. **A probabilistic model of RNA conformational space**. *PLoS Computational Biology*, **5**(6):e1000406, 2009. 14

[86] T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. Johansson, and T. Hamelryck. **Beyond rotamers: a generative, probabilistic model of side chains in proteins**. *BMC Bioinformatics*, **11**(1):306, 2010. 14, 15

[87] Wouter Boomsma. *Protein Structure: probabilisti cmodeling and simulation*. PhD thesis, Department of Biology, University of Copenhage, Denmark, 2008. 15

[88] A.B. Mamonov, D. Bhatt, D.J. Cashman, Y. Ding, and D.M. Zuckerman. **General library-based monte carlo technique enables equilibrium sampling of semi-atomistic protein models**. *The Journal of Physical Chemistry B*, **113**(31):10891–10904, 2009. 15

[89] T. Hamelryck, K. Mardia, and K. Ferkinghoff-Borg. *Bayesian Methods in Structural Bioinformatics*. Springer, New York, 2012. 15

[90] R.H. Swendsen and J.S. Wang. **Nonuniversal critical dynamics in Monte Carlo simulations**. *Physical Review Letters*, **58**(2):86, 1987. 16

[91] U. Wolff. **Collective Monte Carlo updating for spin systems**. *Physical Review Letters*, **62**(4):361, 1989. 16

[92] J. Liu and E. Luijten. **Generalized geometric cluster algorithm for fluid simulation**. *Physical Review E*, **71**(6):066701, 2005. 16

[93] P. Carnevali, G. Tóth, G. Toubassi, and S.N. Meshkat. **Fast protein structure prediction using Monte Carlo simulations with modal moves**. *Journal of the American Chemical Society*, **125**(47):14244, 2003. 16

[94] C. Pangali, M. Rao, and BJ Berne. **On a novel Monte Carlo scheme for simulating water and aqueous solutions**. *Chemical Physics Letters*, **55**(3):413, 1978. 16

[95] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. **Hybrid monte carlo**. *Physics Letters B*, **195**(2):216, 1987. 16

[96] J.A. Izaguirre and S.S. Hampton. **Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules**. *Journal of Computational Physics*, **200**(2):581, 2004. 16

[97] R.H. Swendsen and J.S. Wang. **Replica Monte Carlo simulation of spin-glasses**. *Physical Review Letters*, **57**(21):2607, 1986. 16

[98] B.A. Berg and T. Neuhaus. **Multicanonical algorithms for first order phase transitions**. *Physics Letters B*, **267**(2):249, 1991. 16

[99] K. Hukushima and K. Nemoto. **Exchange Monte Carlo method and application to spin glass simulations**. *Journal of the Physical Society of Japan*, **65**(6):1604, 1996. 16

[100] Y. Levy and J.N. Onuchic. **Water mediation in protein folding and molecular recognition**. *Annual Review of Biophysics and Biomolecular Structure*, **35**:389, 2006. 17

[101] M. Gerstein and C. Chothia. **Packing at the protein-water interface**. *Proceedings of the National Academy of Sciences*, **93**(19):10167, 1996. 17

[102] G. Otting and K. Wuethrich. **Studies of protein hydration in aqueous solution by direct NMR observation of individual protein-bound water molecules**. *Journal of the American Chemical Society*, **111**(5):1871, 1989. 17

[103] Z. Guo, D. Thirumalai, and JD Honeycutt. **Folding kinetics of proteins: A model study**. *The Journal of Chemical Physics*, **97**:525, 1992. 17

[104] M. Mezei, P.J. Fleming, R. Srinivasan, and G.D. Rose. **Polyproline II helix is the preferred conformation for unfolded polyalanine in water**. *Proteins: Structure, Function, and Bioinformatics*, **55**(3):502, 2004. 17

[105] G.A. Papoian, J. Ulander, and P.G. Wolynes. **Role of water mediated interactions in protein-protein recognition landscapes**. *Journal of the American Chemical Society*, **125**(30):9170, 2003. 17

[106] J.R.H. Tame, S.H. Sleigh, A.J. Wilkinson, and J.E. Ladbury. **The role of water in sequence-independent ligand binding by an oligopeptide transporter protein**. *Nature Structural & Molecular Biology*, **3**(12):998, 1996. 17

[107] M. Feig, M. Reiher, and A. Wolf. *Modeling solvent environments*. Wiley Online Library, 2010. 17, 19

[108] W.L. Jorgensen. **Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water**. *Journal of the American Chemical Society*, **103**(2):335, 1981. 17

[109] HJC Berendsen, JR Grigera, and TP Straatsma. **The missing term in effective pair potentials**. *Journal of Physical Chemistry*, **91**(24):6269, 1987. 17

[110] Y.M. Rhee, E.J. Sorin, G. Jayachandran, E. Lindahl, and V.S. Pande. **Simulations of the role of water in the protein-folding mechanism**. *Proceedings of the National Academy of Sciences*, **101**(17):6456, 2004. 17

[111] B. Marten, K. Kim, C. Cortis, R.A. Friesner, R.B. Murphy, M.N. Ringnalda, D. Sitkoff, and B. Honig. **New model for calculation of solvation free energies: correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects**. *The Journal of Physical Chemistry*, **100**(28):11775, 1996. 17

[112] J. Chen, C.L. Brooks, and J. Khandogin. **Recent advances in implicit solvent-based methods for biomolecular simulations**. *Current Opinion in Structural Biology*, **18**(2):140, 2008. 17, 19

[113] T. Lazaridis and M. Karplus. **Effective energy function for proteins in solution**. *Proteins: Structure, Function, and Bioinformatics*, **35**(2):133, 1999. 18, 19

[114] A. Ben-Naim. **Standard thermodynamics of transfer. Uses and misuses**. *The Journal of Physical Chemistry*, **82**(7):792, 1978. 18

[115] D. Eisenberg, A.D. McLachlan, et al. **Solvation energy in protein folding and binding**. *Nature*, **319**(6050):199, 1986. 18

[116] P. Ferrara, J. Apostolakis, and A. Caflisch. **Evaluation of a fast implicit solvent model for molecular dynamics simulations**. *Proteins: Structure, Function, and Bioinformatics*, **46**(1):24, 2002. 18

[117] P.F.W. Stouten, C. Frömmel, H. Nakamura, and C. Sander. **An effective solvation term based on atomic occupancies for use in protein simulations**. *Molecular Simulation*, **10**(2):97, 1993. 18

[118] J. Warwicker and HC Watson. **Calculation of the electric potential in the active site cleft due to α-helix dipoles**. *Journal of Molecular Biology*, **157**(4):671, 1982. 18, 19

[119] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson. **Semianalytical treatment of solvation for molecular mechanics and dynamics**. *Journal of the American Chemical Society*, **112**(16):6127, 1990. 18, 19

[120] D. Qiu, P.S. Shenkin, F.P. Hollinger, and W.C. Still. **The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii**. *The Journal of Physical Chemistry A*, **101**(16):3005, 1997. 18, 19

[121] J. Chen, W. Im, and C.L. Brooks III. **Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field**. *Journal of the American Chemical Society*, **128**(11):3728, 2006. 18, 19

[122] U. Haberthür and A. Caflisch. **FACTS: fast analytical continuum treatment of solvation**. *Journal of Computational Chemistry*, **29**(5):701, 2008. 18, 19

[123] R. Constanciel. **Theoretical basis of the empirical reaction field approximations through continuum model**. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, **69**(5):505, 1986. 18, 19

[124] G.D. Hawkins, C.J. Cramer, and D.G. Truhlar. **Pairwise solute descreening of solute charges from a dielectric medium**. *Chemical Physics Letters*, **246**(1-2):122, 1995. 18, 19

[125] M. Schaefer and M. Karplus. **A comprehensive analytical treatment of continuum electrostatics**. *The Journal of Physical Chemistry*, **100**(5):1578, 1996. 18, 19

[126] A. Ghosh, C.S. Rapp, and R.A. Friesner. **Generalized Born model based on a surface integral formulation**. *The Journal of Physical Chemistry B*, **102**(52):10983, 1998. 18, 19

[127] M.S. Lee, F.R. Salsbury, and C.L. Brooks. **Novel generalized Born methods**. *Journal of Chemical Physics*, **116**(24):10606, 2002. 18, 19

[128] W. Im, M.S. Lee, and C.L. Brooks III. **Generalized born model with a simple smoothing function**. *Journal of Computational Chemistry*, **24**(14):1691, 2003. 18, 19

[129] A. Onufriev, D. Bashford, and D.A. Case. **Exploring protein native states and large-scale conformational changes with a modified generalized born model**. *Proteins: Structure, Function, and Bioinformatics*, **55**(2):383, 2004. 18, 19

[130] E. Gallicchio and R.M. Levy. **AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling**. *Journal of Computational Chemistry*, **25**(4):479, 2004. 18, 19

[131] P.L. Privalov and G.I. Makhatadze. **Contribution of Hydration to Protein Folding Thermodynamics:: II. The Entropy and Gibbs Energy of Hydration**. *Journal of Molecular Biology*, **232**(2):660, 1993. 19

[132] E. Neria, S. Fischer, and M. Karplus. **Simulation of activation free energies in molecular systems**. *The Journal of Chemical Physics*, **105**(5):1902, 1996. 20, 21

[133] D. Shirvanyants, F. Ding, D. Tsao, S. Ramanchandran, and N.V. Dokholyan. **DMD: An Efficient and Versatile Simulation Method for Fine Protein Characterization**. *The Journal of Physical Chemistry B*, 2012. 20

[134] A. Cavalli, X. Salvatella, C.M. Dobson, and M. Vendruscolo. **Protein structure determination from NMR chemical shifts**. *Proceedings of the National Academy of Sciences*, **104**(23):9615, 2007. 20

[135] K.W. Kaufmann, G.H. Lemmon, S.L. DeLuca, J.H. Sheehan, and J. Meiler. **Practically useful: what the Rosetta protein modeling suite can do for you**. *Biochemistry*, **49**(14):2987, 2010. 20

[136] RL McGreevy and L. Pusztai. **Reverse Monte Carlo simulation: a new technique for the determination of disordered structures**. *Molecular Simulation*, **1**(6):359, 1988. 21

[137] F. Ercolessi and J.B. Adams. **Interatomic potentials from first-principles calculations: the force-matching method**. *Europhysics Letters*, **26**(8):583, 1994. 21

[138] S. Izvekov, M. Parrinello, C.J. Burnham, and G.A. Voth. **Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching**. *The Journal of Chemical Physics*, **120**(23):10896, 2004. 21

[139] S. Izvekov and G.A. Voth. **A multiscale coarse-graining method for biomolecular systems**. *The Journal of Physical Chemistry B*, **109**(7):2469, 2005. 21

[140] J. Zhou, I.F. Thorpe, S. Izvekov, and G.A. Voth. **Coarse-grained peptide modeling using a systematic multiscale approach**. *Biophysical journal*, **92**(12):4289, 2007. 21

[141] M.S. Shell. **The relative entropy is fundamental to multiscale and inverse thermodynamic problems**. *The Journal of Chemical Physics*, **129**(14):144108, 2008. 21

[142] R. B. Best, X. Zhu, J. Shim, P. Lopes, J. Mittal, M Feig, and MacKerell A.D. **Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ,$\psi$and side-chain χ_1 and χ_2 dihedral angles.** *Unpublished manuscript*, 2012. 22

[143] A. Chaimovich and M.S. Shell. **Relative entropy as a universal metric for multiscale errors**. *Physical Review E*, **81**(6):060104, 2010. 23

[144] A. Chaimovich and M.S. Shell. **Coarse-graining errors and numerical optimization using a relative entropy framework**. *The Journal of Chemical Physics*, **134**(9):094112, 2011. 23

[145] S.P. Carmichael and M.S. Shell. **A New Multiscale Algorithm and its Application to Coarse-Grained Peptide Models for Self-Assembly**. *The Journal of Physical Chemistry B*, 2012. 23

[146] R. Zhou. **Free energy landscape of protein folding in water: explicit vs. implicit solvent**. *Proteins: Structure, Function, and Bioinformatics*, **53**(2):148, 2003. 124

[147] M.S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A.L. Beberg, D.L. Ensign, C.M. Bruns, and V.S. Pande. **Accelerating molecular dynamic simulation on graphics processing units**. *Journal of Computational Chemistry*, **30**(6):864, 2009. 124

[148] S.A. Hassan, F. Guarnieri, and E.L. Mehler. **A general treatment of solvent effects based on screened Coulomb potentials**. *The Journal of Physical Chemistry B*, **104**(27):6478, 2000. 124

[149] K. Stovgaard, C. Andreetta, J. Ferkinghoff-Borg, and T. Hamelryck. **Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models**. *BMC Bioinformatics*, **11**(1):429, 2010. 124

[150] A. Rosato, J.M. Aramini, C. Arrowsmith, A. Bagaria, D. Baker, A. Cavalli, J.F. Doreleijers, A. Eletsky, A. Giachetti, P. Guerry, et al. **Blind testing of routine, fully automated determination of protein structures from NMR data**. *Structure*, **20**(2):2270, 2012. 124

[151] J. Hu, A. Ma, and A.R. Dinner. **Monte Carlo simulations of biomolecules: The MC module in CHARMM**. *Journal of Computational Chemistry*, **27**(2):203, 2006. 124

4

Appendix A

Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized

This article presents the reference ratio method for combining two probability distributions describing respectively local and non-local features of biomolecular structure. A theoretical formulation for the problem is given, and two applications of the method to protein structure determination are presented as explicative examples.

Although I was not involved in the theoretical discussions leading to the developement of the reference ratio method, I contributed with the sampling algorithms used in the study.

OPEN ACCESS Freely available online

PLoS one

Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized

Thomas Hamelryck[1*9], Mikael Borg[19], Martin Paluszewski[19], Jonas Paulsen[1], Jes Frellsen[1], Christian Andreetta[1], Wouter Boomsma[2,3], Sandro Bottaro[2], Jesper Ferkinghoff-Borg[2*]

1 Bioinformatics Center, Department of Biology, University of Copenhagen, Copenhagen, Denmark, **2** Biomedical Engineering, Technical University of Denmark (DTU) Elektro, Technical University of Denmark, Lyngby, Denmark, **3** Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Abstract

Understanding protein structure is of crucial importance in science, medicine and biotechnology. For about two decades, knowledge-based potentials based on pairwise distances – so-called "potentials of mean force" (PMFs) – have been center stage in the prediction and design of protein structure and the simulation of protein folding. However, the validity, scope and limitations of these potentials are still vigorously debated and disputed, and the optimal choice of the reference state – a necessary component of these potentials – is an unsolved problem. PMFs are loosely justified by analogy to the reversible work theorem in statistical physics, or by a statistical argument based on a likelihood function. Both justifications are insightful but leave many questions unanswered. Here, we show for the first time that PMFs can be seen as approximations to quantities that do have a rigorous probabilistic justification: they naturally arise when probability distributions over different features of proteins need to be combined. We call these quantities "reference ratio distributions" deriving from the application of the "reference ratio method." This new view is not only of theoretical relevance but leads to many insights that are of direct practical use: the reference state is uniquely defined and does not require external physical insights; the approach can be generalized beyond pairwise distances to arbitrary features of protein structure; and it becomes clear for which purposes the use of these quantities is justified. We illustrate these insights with two applications, involving the radius of gyration and hydrogen bonding. In the latter case, we also show how the reference ratio method can be iteratively applied to sculpt an energy funnel. Our results considerably increase the understanding and scope of energy functions derived from known biomolecular structures.

Citation: Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, et al. (2010) Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. PLoS ONE 5(11): e13714. doi:10.1371/journal.pone.0013714

Editor: Darren R. Flower, University of Oxford, United Kingdom

Received July 7, 2010; **Accepted** October 4, 2010; **Published** November 10, 2010

Funding: The authors acknowledge funding by the Danish Program Commission on Nanoscience, Biotechnology and IT (NaBiIT, project: Simulating proteins on a millisecond time-scale, 2106-06-0009), the Danish Research Council for Technology and Production Sciences (FTP, project: Protein structure ensembles from mathematical models, 274-09-0184) and the Danish Council for Independent Research (FNU, project: A Bayesian approach to protein structure determination, 272-08-0315). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: thamelry@binf.ku.dk (TH); jfb@elektro.dtu.dk (JFB)

9 These authors contributed equally to this work.

Introduction

Methods for protein structure prediction, simulation and design rely on an energy function that represents the protein's free energy landscape; a protein's native state typically corresponds to the state with minimum free energy [1]. So-called knowledge based potentials (KBP) are parametrized functions for free energy calculations that are commonly used for modeling protein structures [2,3]. These potentials are obtained from databases of known protein structures and lie at the heart of some of the best protein structure prediction methods. The use of KBPs originates from the work of Tanaka and Scheraga [4] who were the first to extract effective interactions from the frequency of contacts in X-ray structures of native proteins. Miyazawa and Jernigan formalized the theory for contact interactions by means of the quasi-chemical approximation [5,6].

Many different approaches for developing KBPs exist, but the most successful methods to date build upon a seminal paper by Sippl – published two decades ago – which introduced KBPs based on probability distributions of pairwise distances in proteins and reference states [7]. These KBPs were called "potentials of mean force", and seen as approximations of free energy functions. Sippl's work was inspired by the statistical physics of liquids, where a "potential of mean force" has a very precise and undisputed definition and meaning [8,9]. However, the validity of the application to biological macromolecules is vigorously disputed in the literature [2,10–17]. Nonetheless, PMFs are widely used with considerable success; not only for protein structure prediction [3,18,19], but also for quality assessment and identification of errors [20–22], fold recognition and threading [23,24], molecular dynamics [24], protein-ligand interactions [16,25], protein design and engineering [26,27], and the prediction of binding affinity [17,28]. In this article, the abbreviation "PMF" will refer to the pairwise distance dependent KBPs following Sippl [7], and the generalization that we introduce in this article; we will write "potentials of mean force" in full when we refer to the real, physically valid potentials as used in liquid systems [9,13,29]. At the end of the article, we will propose a new name for these statistical quantities, to set them apart from true potentials of mean force with a firm physical basis.

Despite the progress in methodology and theory, and the dramatic increase in the number of experimentally determined protein structures, the accuracy of the energy functions still remains the main obstacle to accurate protein structure prediction [22,30,31]. Recently, several groups demonstrated that it is the quality of the coarse grained energy functions [18], rather than inadequate sampling, that impairs the successful prediction of the native state [30,31]. The insights presented in this article point towards a new, theoretically well-founded way to construct and refine energy functions, and thus address a timely problem.

We start with an informal outline of the general ideas presented in this article, and then analyze two notable attempts in the literature to justify PMFs. We point out their shortcomings, and subsequently present a rigorous probabilistic explanation of the strengths and shortcomings of traditional pairwise distance PMFs. This explanation sheds a surprising new light on the nature of the reference state, and allows the generalization of PMFs beyond pairwise distances in a statistically valid way. Finally, we demonstrate our method in two applications involving protein compactness and hydrogen bonding. In the latter case, we also show that PMFs can be iteratively optimized, thereby effectively sculpting an energy funnel [24,32–36].

Results and Discussion

Overview

In order to emphasize the practical implications of the theoretical insights that we present here, we start with a very concrete example that illustrates the essential concepts (see Fig. 1). Currently, protein structure prediction methods often make use of fragment libraries: collections of short fragments derived from known protein structures in the Protein Data Bank (PDB). By assembling a suitable set of fragments, one obtains conformations that are protein-like on a local length scale. That is, these conformations typically lack non-local features that characterize real proteins, such as a well-packed hydrophobic core or an extensive hydrogen bond network. Such aspects of protein structure are not, or only partly, captured by fragment libraries.

Formally, a fragment library specifies a probability distribution $Q(X)$, where X is for example a vector of dihedral angles. In order to obtain conformations that also possess the desired non-local features, $Q(X)$ needs to be complemented with another probability distribution $P(Y)$, with Y being for example a vector of pairwise distances, the radius of gyration, the hydrogen bonding network, or any combination of non-local features. Typically, Y is a deterministic function of X; we use the notation $Y(X)$ when necessary.

For the sake of argument, we will focus on the radius of gyration (r_g) at this point; in this case $Y(X)$ becomes $r_g(X)$. We assume that a suitable $P(r_g)$ was derived from the set of known protein structures; without loss of generality, we leave out the dependency on the amino acid sequence for simplicity. The problem that we address in this article can be illustrated with the following question: how can we combine $P(r_g)$ and $Q(X)$ in a rigorous, meaningful way? In other words, we want to use the fragment library to sample conformations whose radii of gyration r_g are distributed according to $P(r_g)$. These conformations should display a realistic *local* structure as well, reflecting the use of the fragment library. Simply multiplying $P(r_g(X))$ and $Q(X)$ does not lead to the desired result, as X and R_g are not independent; the resulting conformations will not be distributed according to $P(r_g)$.

The solution is given in Fig. 1; it involves the probability distribution $Q_R(r_g)$, the probability distribution over the radius of gyration for conformations sampled solely from the fragment library. The subscript R stands for *reference state* as will be explained below. The solution generates conformations whose radii of gyration are distributed according to $P(r_g)$. The influence of $Q(X)$ is apparent in the fact that for conformations with a given r_g, their local structure X will be distributed according to $Q(X|r_g)$. The

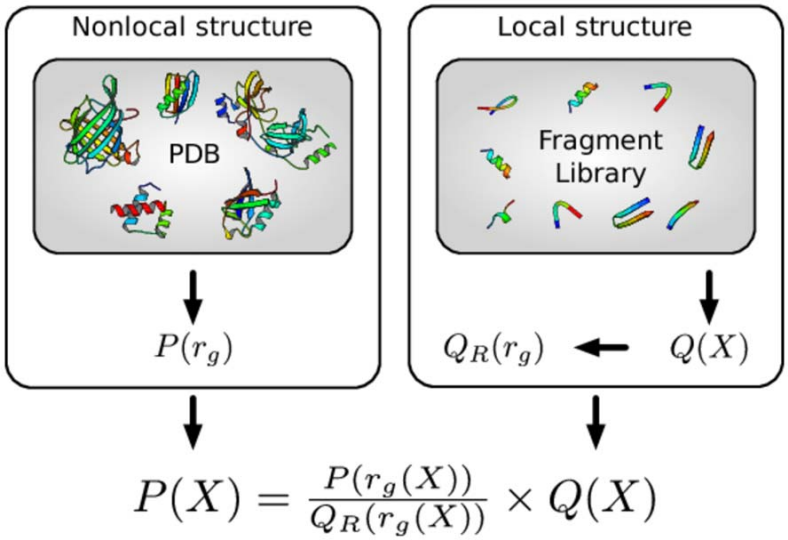


Figure 1. Illustration of the central idea presented in this article. In this example, the goal is to sample conformations with a given distribution $P(r_g)$ for the radius of gyration r_g, and a plausible local structure. $P(r_g)$ could, for example, be derived from known structures in the Protein Data Bank (PDB, left box). $Q(X)$ is a probability distribution over local structure X, typically embodied in fragment library (right box). In order to combine $Q(X)$ and $P(r_g)$ in a meaningful way (see text), the two distributions are multiplied and divided by $Q_R(r_g)$ (formula at the bottom); $Q_R(r_g)$ is the probability distribution over the radius of gyration for conformations sampled solely from the fragment library (that is, $Q(X)$). The probability distribution $P(X)$ will generate conformations with plausible local structures (due to $Q(X)$), while their radii of gyration will be distributed according to $P(r_g)$, as desired. This simple idea lies at the theoretical heart of the PMF expressions used in protein structure prediction.
doi:10.1371/journal.pone.0013714.g001

latter distribution has a clear interpretation: it corresponds to sampling an infinite amount of conformations from a fragment library, and retaining only those with the desired r_g. Note that even if we chose the uniform distribution for $Q(X)$, the resulting $Q_R(r_g)$ will *not* (necessarily) be uniform.

Intuitively, $P(r_g)$ provides correct information about the radius of gyration, but no information about local structure; $Q(X)$ provides approximately correct information about the structure of proteins on a local length scale, but is incorrect on a global scale (leading to an incorrect probability distribution for the radius of gyration); finally, the formula shown in Fig. 1 merges these two complementary sources of information together. Another viewpoint is that $P(r_g)$ and $Q(r_g)$ are used to correct the shortcomings of $Q(X)$. This construction is statistically rigorous, provided that $P(r_g)$ and $Q(X)$ are proper probability distributions.

After this illustrative example, we now review the use of PMFs in protein structure prediction, and discuss how PMFs can be understood and generalized in the theoretical framework that we briefly outlined here.

Pairwise PMFs for protein structure prediction

Many textbooks present PMFs as a simple consequence of the Boltzmann distribution, as applied to pairwise distances between amino acids. This distribution, applied to a specific pair of amino acids, is given by:

$$P(r) = \frac{1}{Z} e^{-\frac{F(r)}{kT}}$$

where r is the distance, k is Boltzmann's constant, T is the temperature and Z is the partition function, with $Z = \int e^{-\frac{F(r)}{kT}} dr$. The quantity $F(r)$ is the free energy assigned to the pairwise system. Simple rearrangement results in the *inverse Boltzmann formula*, which expresses the free energy $F(r)$ as a function of $P(r)$:

$$F(r) = -kT \ln P(r) - kT \ln Z$$

To construct a PMF, one then introduces a so-called *reference state* with a corresponding distribution Q_R and partition function Z_R, and calculates the following free energy difference:

$$\Delta F(r) = -kT \ln \frac{P(r)}{Q_R(r)} - kT \ln \frac{Z}{Z_R} \quad (1)$$

The reference state typically results from a hypothetical system in which the specific interactions between the amino acids are absent [7]. The second term involving Z and Z_R can be ignored, as it is a constant.

In practice, $P(r)$ is estimated from the database of known protein structures, while $Q_R(r)$ typically results from calculations or simulations. For example, $P(r)$ could be the conditional probability of finding the $C\beta$ atoms of a valine and a serine at a given distance r from each other, giving rise to the free energy difference ΔF. The total free energy difference of a protein, ΔF_{TOT}, is then claimed to be the sum of all the pairwise free energies:

$$\Delta F_{\mathrm{TOT}} = \sum_{i<j} \Delta F(r_{ij}|a_i,a_j) \quad (2)$$

$$= -kT \sum_{i<j} \ln \frac{P(r_{ij}|a_i,a_j)}{Q_R(r_{ij}|a_i,a_j)} \quad (3)$$

where the sum runs over all amino acid pairs a_i,a_j (with $i<j$) and r_{ij} is their corresponding distance. It should be noted that in many studies Q_R does not depend on the amino acid sequence [11].

Intuitively, it is clear that a low free energy difference indicates that the set of distances in a structure is more likely in proteins than in the reference state. However, the physical meaning of these PMFs have been widely disputed since their introduction [2,12–15]. Indeed, why is it at all necessary to subtract a reference state energy? What is the optimal reference state? Can PMFs be generalized and justified beyond pairwise distances, and if so, how? Before we discuss and clarify these issues, we discuss two qualitative justifications that were previously reported in the literature: the first based on a physical analogy, and the second using a statistical argument.

PMFs from the reversible work theorem

The first, qualitative justification of PMFs is due to Sippl, and based on an analogy with the statistical physics of liquids [37]. For liquids [8,9,13,14,37], the potential of mean force is related to the *pair correlation function* $g(r)$, which is given by:

$$g(r) = \frac{P(r)}{Q_R(r)}$$

where $P(r)$ and $Q_R(r)$ are the respective probabilities of finding two particles at a distance r from each other in the liquid and in the reference state. For liquids, the reference state is clearly defined; it corresponds to the ideal gas, consisting of non-interacting particles. The two-particle potential of mean force $W(r)$ is related to $g(r)$ by:

$$W(r) = -kT \log g(r) = -kT \log \frac{P(r)}{Q_R(r)} \quad (4)$$

According to the *reversible work theorem*, the two-particle potential of mean force $W(r)$ is the reversible work required to bring two particles in the liquid from infinite separation to a distance r from each other [8,9].

Sippl justified the use of PMFs – a few years after he introduced them for use in protein structure prediction [7] – by appealing to the analogy with the reversible work theorem for liquids [37]. For liquids, $g(r)$ can be experimentally measured using small angle X-ray scattering; for proteins, $P(r)$ is obtained from the set of known protein structures, as explained in the previous section. The analogy described above might provide some physical insight, but, as Ben-Naim writes in a seminal publication [13]: "the quantities, referred to as 'statistical potentials,' 'structure based potentials,' or 'pair potentials of mean force', as derived from the protein data bank, are neither 'potentials' nor 'potentials of mean force,' in the ordinary sense as used in the literature on liquids and solutions."

Another issue is that the analogy does not specify a suitable reference state for proteins. This is also reflected in the literature on statistical potentials; the construction of a suitable reference state continues to be an active research topic [3,22,38–41]. In the next section, we discuss a second, more recent justification that is based on probabilistic reasoning.

PMFs from likelihoods

Baker and co-workers [18] justified PMFs from a Bayesian point of view and used these insights in the construction of the coarse

grained ROSETTA energy function; Samudrala and Moult used similar reasoning for the RAPDF potential [42]. According to Bayesian probability calculus, the conditional probability $P(X|A)$ of a structure X, given the amino acid sequence A, can be written as:

$$P(X|A) = \frac{P(A|X)P(X)}{P(A)} \propto P(A|X)P(X)$$

$P(X|A)$ is proportional to the product of the likelihood $P(A|X)$ times the prior $P(X)$. By assuming that the likelihood can be approximated as a product of pairwise probabilities, and applying Bayes' theorem, the likelihood can be written as:

$$P(A|X) \approx \prod_{i<j} P(a_i,a_j|r_{ij}) \propto \prod_{i<j} \frac{P(r_{ij}|a_i,a_j)}{P(r_{ij})} \qquad (5)$$

where the product runs over all amino acid pairs a_i,a_j (with $i<j$), and r_{ij} is the distance between amino acids i and j. Obviously, the negative of the logarithm of expression (5) has the same functional form as the classic pairwise distance PMFs, with the denominator playing the role of the reference state in Eq. 1. The merit of this explanation is the qualitative demonstration that the functional form of a PMF can be obtained from probabilistic reasoning. Although this view is insightful – it rightfully drew the attention to the application of Bayesian methods to protein structure prediction – there is a more quantitative explanation, which does not rely on the incorrect assumption of pairwise decomposability [12–14,43], and leads to a different, *quantitative* conclusion regarding the nature of the reference state. This explanation is given in the next section.

A general statistical justification for PMFs

Expressions that resemble PMFs naturally result from the application of probability theory to solve a fundamental problem that arises in protein structure prediction: how to improve an imperfect probability distribution $Q(X)$ over a first variable X using a probability distribution $P(Y)$ over a second variable Y (see Fig. 2, Fig. 1 and Materials and Methods). We assume that Y is a deterministic function of X; we write $Y(X)$ when necessary. In that case, X and Y are called *fine* and *coarse grained variables*, respectively. When Y is a function of X, the probability distribution $Q(X)$ automatically implies a probability distribufotion $Q(X,Y(X))$. This distribution has some unusual properties: $Q(X,Y(X)) = Q(X)$; and if $Y' \neq Y(X)$, it follows that $Q(X,Y') = 0$.

Typically, X represents *local* features of protein structure (such as backbone dihedral angles), while Y represents *nonlocal* features (such as hydrogen bonding, compactness or pairwise distances). However, the same reasoning also applies to other cases; for example, $P(Y)$ could represent information coming from experimental data, and $Q(X)$ could be embodied in an empirical force field as used in molecular mechanics [2,44] (see Fig. 2).

Typically, the distribution $Q(X)$ in itself is not sufficient for protein structure prediction: it does not consider important nonlocal features such as hydrogen bonding, compactness or favorable amino acid interactions. As a result, $Q(X)$ is incorrect with respect to Y, and needs to be supplemented with a probability distribution $P(Y)$ that provides additional information. By construction, $P(Y)$ is assumed to be correct (or at least useful).

The above situation arises naturally in protein structure prediction. For example, $P(Y)$ could be a probability distribution over the radius of gyration, hydrogen bond geometry or the set of pairwise distances, and $Q(X)$ could be a fragment library [18] or a

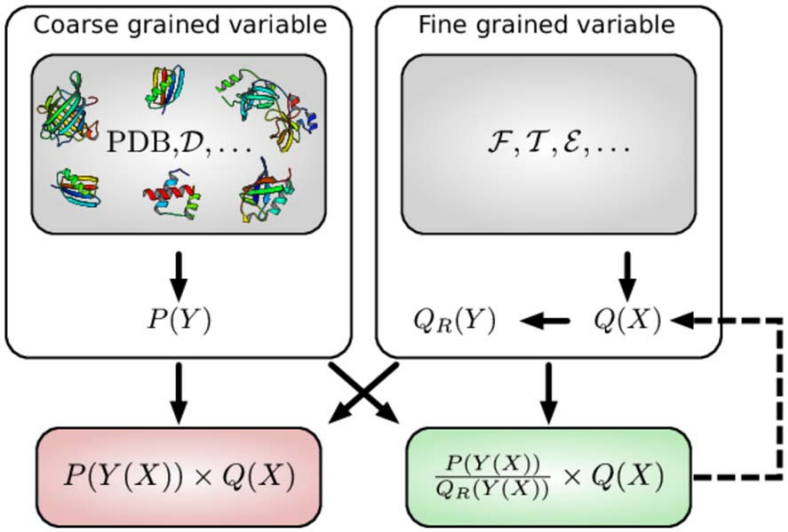


Figure 2. General statistical justification of PMFs. The goal is to combine a distribution $Q(X)$ over a fine grained variable X (top right), with a probability distribution $P(Y)$ over a coarse grained variable $Y(X)$ (top left). $Q(X)$ could be, for example, embodied in a fragment library ($\mathcal{F}$), a probabilistic model of local structure ($\mathcal{T}$) or an energy function ($\mathcal{E}$); Y could be, for example, the radius of gyration, the hydrogen bond network, or the set of pairwise distances. $P(Y)$ usually reflects the distribution of Y in known protein structures (PDB), but could also stem from experimental data ($\mathcal{D}$). Sampling from $Q(X)$ results in a distribution $Q_R(Y)$ that differs from $P(Y)$. Multiplying $P(Y)$ and $Q(X)$ does not result in the desired distribution for Y either (red box); the correct result requires dividing out the signal with respect to Y due to $Q(X)$ (green box). The *reference* distribution $Q_R(Y)$ in the denominator corresponds to the contribution of the reference state in a PMF. If $Q_R(Y)$ is only approximately known, the method can be applied iteratively (dashed arrow). In that case, one attempts to iteratively sculpt an energy funnel. The procedure is statistically rigorous provided $Q(X)$ and $P(Y)$ are proper probability distributions; this is usually not the case for conventional pairwise distance PMFs.
doi:10.1371/journal.pone.0013714.g002

probabilistic model of local structure [45]. In Fig. 1, we used the example of a distribution over the radius of gyration for $P(Y)$ and a fragment library for $Q(X)$. Obviously, sampling from a fragment library and retaining structures with the desired nonlocal structure (radius of gyration, hydrogen bonding, etc.) is in principle possible, but in practice extremely inefficient.

How can $Q(X)$ be combined with $P(Y)$ in a meaningful way? As mentioned previously, simply multiplying the two distributions – resulting in $P(Y(X))Q(X)$ – does not lead to the desired result as the two variables are obviously not independent. The correct solution follows from simple statistical considerations (see Materials and Methods), and is given by the following expression:

$$P(X) = \frac{P(Y(X))}{Q_R(Y(X))} Q(X) \tag{6}$$

We use the notation $P(X)$, as this distribution implies the desired distribution $P(Y)$ for $Y(X)$. The distribution $Q_R(Y)$ in the denominator is the probability distribution that is implied by $Q(X)$ over the coarse grained variable Y. Conceptually, dividing by $Q_R(Y)$ takes care of the signal in $Q(X)$ with respect to the coarse grained variable Y. The ratio in this expression corresponds to the probabilistic formulation of a PMF, and $Q_R(Y)$ corresponds to the reference state (see Materials and Methods).

In practice, $Q(X)$ is typically not evaluated directly, but brought in through conformational Monte Carlo sampling (see Materials and Methods); often sampling is based on a fragment library [18,46], although other methods are possible, including sampling from a probabilistic model [45,47,48] or a suitable energy function [2,44]. The ratio $P(Y)/Q_R(Y)$, which corresponds to the probabilistic formulation of a PMF, also naturally arises in the Markov chain Monte Carlo (MCMC) procedure (see Materials and Methods). An important insight is that, in this case, the conformational sampling method uniquely defines the reference state. Thus, in the case of a fragment library, the reference distribution $Q_R(Y)$ is the probability distribution over Y that is obtained by sampling conformations solely using the fragment library.

As the method we have introduced here invariably relies on the ratio of two probability distributions – one regarding protein structure and the other regarding a well-defined reference state – we refer to it as the *reference ratio method*. In the next section, we show that the standard pairwise distance PMFs can be seen as an approximation of the reference ratio method.

Pairwise distance PMFs explained

In this section, we apply the reference ratio method to the standard, pairwise distance case. In the classic PMF approach, one considers the vector of pairwise distances R between the amino acids. In this case, it is usually assumed that we can write

$$P(R|A) \propto \prod_{i<j} P(r_{ij}|a_i,a_j) \tag{7}$$

where the product runs over all amino acid pairs a_i,a_j (with $i<j$), and r_{ij} is their matching distance. Clearly, the assumption that the joint probability can be written as a product of pairwise probabilities is not justified [12,13,43], but in practice this assumption often provides useful results [22]. In order to obtain protein-like conformations, $P(R|A)$ needs to be combined with an appropriate probability distribution $Q(X|A)$ that addresses the local features of the polypeptide chain. Applying Eq. 6 to this case results in the following expression:

$$P(X|A) \propto \frac{\prod_{i<j} P(r_{ij}|a_i,a_j)}{\prod_{i<j} Q_R(r_{ij}|a_i,a_j)} Q(X|A)$$

where the denominator $Q_R(\cdot)$ is the probability distribution over the pairwise distances as induced by the distribution $Q(X|A)$. The ratio in this expression corresponds to the probabilistic expression of a PMF. The reference state is thus determined by $Q(X|A)$: it reflects the probability of generating a set of pairwise distances using local structure information alone. Obviously, as $Q(X|A)$ is conditional upon the amino acid sequence A, the reference state becomes sequence dependent as well.

We again emphasize that the assumption of pairwise decomposability in Eq. 7 is incorrect [12–14,43]. Therefore, the application of the reference ratio method results in a useful approximation, at best. As a result, the optimal definition of the reference state also needs to compensate for the errors implied by the invalid assumption. As is it well established that distance dependent PMFs perform well with a suitable definition of the reference state [3,22,38–40], and the incorrect pairwise decomposability assumption impairs a rigorous statistical analysis, we do not discuss this type of PMFs further. Indeed, for pairwise distance PMFs, the main challenge lies in developing better probabilistic models of sets of pairwise distances [49].

The pairwise distance PMFs currently used in protein structure prediction are thus not statistically rigorous, because they do not make use of a proper joint probability distribution over the pairwise distances, which are strongly intercorrelated due to the connectivity of molecules. A rigorous application of the reference ratio method would require the construction of a proper joint probability distribution over pairwise distances. This is certainly possible in principle, but currently, as far as we know, a challenging open problem and beyond the scope of this article. However, we have clarified that the idea of using a reference state is correct and valid, and that this state has a very precise definition. Therefore, in the next two sections, we show instead how statistically valid quantities, similar to PMFs, can be obtained for very different coarse grained variables.

A generalized PMF: radius of gyration

As a first application of the reference ratio method, we consider the task of sampling protein conformations with a given probability distribution $P(r_g)$ for the radius of gyration r_g. For $P(r_g)$, we chose a Gaussian distribution with mean $\mu = 22$ Å and standard deviation $\sigma = 2$ Å. This choice is completely arbitrary; it simply serves to illustrate that the reference ratio method allows imposing an exact probability distribution over a certain feature of interest. Applying Eq. 6 results in:

$$P(X|A) = \frac{P(r_g(X))}{Q_R(r_g(X)|A)} Q(X|A) \tag{8}$$

For $Q(X|A)$, we used TorusDBN – a graphical model that allows sampling of plausible backbone angles [45] – and sampled conditional on the amino acid sequence A of ubiquitin (see Materials and Methods). $Q_R(r_g|A)$ is the probability distribution of the radius of gyration for structures sampled solely from TorusDBN, which was determined using generalized multihistogram MCMC sampling (see Materials and Methods).

In Fig. 3, we contrast sampling from Eq. 8 with sampling from $P(r_g(X))Q(X|A)$. In the latter case, the reference state is not properly taken into account, which results in a significant shift towards higher radii of gyration. In contrast, the distribution of r_g for the correct distribution $P(X)$, given by Eq. 8, is indistinguish-

able from the target distribution. This qualitative result is confirmed by the Kullback-Leibler divergence [50] – a natural distance measure for probability distributions expressed in bits – between the target distribution and the resulting marginal distributions of r_g. Adding $Q_R(r_g(X)|A)$ to the denominator diminishes the distance from 0.08 to 0.001 bits. For this particular PMF, the effect of using the correct reference state is significant, but relatively modest; in the next section, we discuss an application where its effect is much more pronounced.

Iterative optimization of PMFs: hydrogen bonding

Here, we demonstrate that PMFs can be optimized iteratively, which is particularly useful if the reference probability distribution $Q_R(Y|A)$ is difficult to estimate. We illustrate the method with a target distribution that models the hydrogen bonding network using a multinomial distribution.

We describe the hydrogen bonding network (H) with eight integers (for details, see Materials and Methods). Three integers (n_α,n_β,n_c) represent the number of residues that do not partake in hydrogen bonds in α-helices, β-sheets and coils, respectively. The five remaining integers ($n_{\alpha\alpha}$,$n_{\beta\beta}$,n_{cc},$n_{\alpha c}$,$n_{\beta c}$) represent the number of hydrogen bonds within α-helices, within β-strands, within coils, between α-helices and coils, and between β-strands and coils, respectively.

As target distribution $P(H)$ over these eight integers, we chose a multinomial distribution whose parameters were derived from the native structure of protein G (see Materials and Methods). $P(H)$ provides information, regarding protein G, on the number of hydrogen bonds and the secondary structure elements involved, but does not specify *where* the hydrogen bonds or secondary elements occur. As in the previous section, we use TorusDBN as the sampling distribution $Q(X|A)$; we sample backbone angles conditional on the amino acid sequence A of protein G. Native secondary structure information was *not* used in sampling from TorusDBN.

The reference distribution $Q_R(H|A)$, due to TorusDBN, is very difficult to estimate correctly for several reasons: its shape is unknown and presumably complex; its dimensionality is high; and the data is very sparse with respect to β-sheet content. Therefore, $Q_R(H|A)$ can only be approximated, which results in a suboptimal PMF. A key insight is that one can apply the method iteratively until a satisfactory PMF is obtained (see Fig. 2, dashed line). In each iteration, the (complex) reference distribution is approximated using a simple probability distribution; we illustrate the method by using a multinomial distribution, whose parameters are estimated by maximum likelihood estimation in each iteration, using the conformations generated in the previous iteration. In the first iteration, we simply set the reference distribution equal to the uniform distribution.

Formally, the procedure works as follows. In iteration $i+1$, the distribution $P_i(H|A)$ is improved using the samples generated in iteration i:

$$P_{i+1}(X|A) = \frac{P(H(X))}{P_{R,i}(H(X)|A)} P_i(X|A) \qquad (9)$$

where $P_{R,i}(H|A)$ is the reference distribution estimated from the samples generated in the i-th iteration, $P_0(X) = Q(X|A)$ stems from TorusDBN, and $P_{R,0}(H|A)$ is the uniform distribution. After each iteration, the set of samples is enriched in hydrogen bonds, and the reference distribution $P_{R,i}(H|A)$ can be progressively estimated more precisely. Note that in the first iteration, we simply use the product of the target and the sampling distribution; no reference state is involved.

Fig. 4 shows the evolution of the fractions versus the iteration number for the eight hydrogen bond categories; the structures with minimum energy for all six iterations are shown in Fig. 5. In the

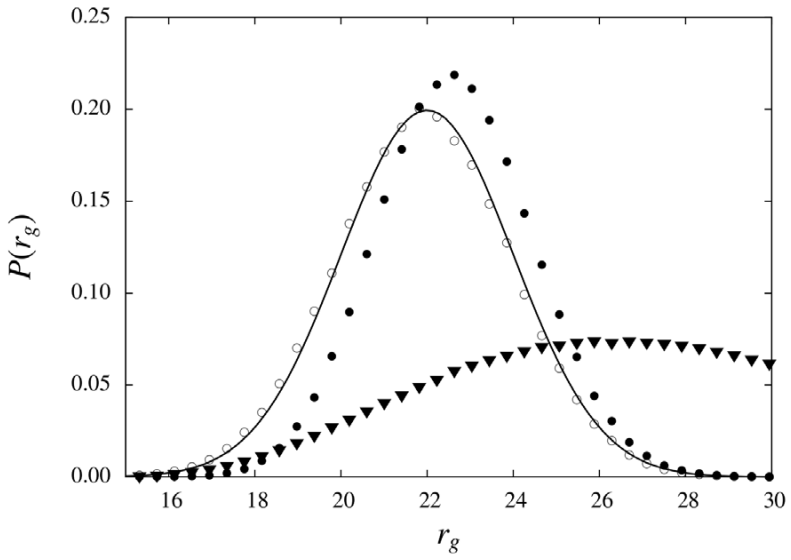


Figure 3. A PMF based on the radius of gyration. The goal is to adapt a distribution $Q(X|A)$ – which allows sampling of local structures – such that a given target distribution $P(r_g)$ is obtained. For A, we used the amino acid sequence of ubiquitin. Sampling from $Q(X|A)$ alone results in a distribution with an average r_g of about 27 A (triangles). Sampling using the correct expression (open circles), given by Eq. 8, results in a distribution that coincides with the target distribution (solid line). Not taking the reference state into account results in a significant shift towards higher r_g (black circles).
doi:10.1371/journal.pone.0013714.g003

first iteration, the structure with minimum energy (highest probability) consists of a single α-helix; β-sheets are entirely absent (see Fig. 5, structure 1). Already in the second iteration, β-strands start to pair, and in the third and higher iterations complete sheets are readily formed. The iterative optimization of the PMF quickly leads to a dramatic enrichment in β-sheet structures, as desired, and the fractions of the eight categories become very close to the native values (Fig. 4).

Conclusions

The strengths and weaknesses of PMFs can be rigorously explained based on simple probabilistic considerations, which leads to some surprising new insights of direct practical relevance. First, we have made clear that PMFs naturally arise when two probability distributions need to be combined in a meaningful way. One of these distributions typically addresses local structure, and its contribution often arises from conformational sampling. Each conformational sampling method thus requires its own reference state and corresponding reference distribution; this is likely the main reason behind the large number of different reference states reported in the literature [3,22,38–41]. If the sampling method is conditional upon the amino acid sequence, the reference state necessarily also depends on the amino acid sequence.

Second, conventional applications of pairwise distance PMFs usually lack two necessary features to make them fully rigorous: the use of a proper probability distribution over pairwise distances in proteins for $P(Y|A)$, and the recognition that the reference state is rigorously defined by the conformational sampling scheme used, that is, $Q(X|A)$. Usually, the reference state is derived from external physical considerations [11,51].

Third, PMFs are not tied to pairwise distances, but generalize to any coarse grained variable. Attempts to develop similar quantities that, for example, consider solvent exposure [52,53], relative side chain orientations [54], backbone dihedral angles [55,56] or hydrogen bonds [37] are thus, in principle, entirely justified. Hence, our probabilistic interpretation opens up a wide range of possibilities for advanced, well-justified energy functions based on sound probabilistic reasoning; the main challenge is to develop proper probabilistic models of the features of interest and the estimation of their parameters [49,57]. Strikingly, the example applications involving radius of gyration and hydrogen bonding that we presented in this article *are* statistically valid and rigorous, in contrast to the traditional pairwise distance PMFs.

Finally, our results reveal a straightforward way to optimize PMFs. Often, it is difficult to estimate the probability distribution that describes the reference state. In that case, one can start with an approximate PMF, and apply the method iteratively. In each iteration, a new reference state is estimated, with a matching probability distribution. In that way, one iteratively attempts to sculpt an energy funnel [24,32–36]. We illustrated this approach with a probabilistic model of the hydrogen bond network. Although iterative application of the inverse Boltzmann formula has been described before [24,35,58,59], its theoretical justification, optimal definition of the reference state and scope remained unclear.

As the traditional pairwise distance PMFs used in protein structure prediction arise from the imperfect application of a statistically valid and rigorous procedure with a much wider scope, we consider it highly desirable that the name "potential of mean force" should be reserved for true, physically valid quantities [13]. Because the statistical quantities we discussed invariably rely on the use of a ratio of two probability distributions, one concerning protein structure and the other concerning the (now well defined) reference state, we suggest the name "reference ratio distribution" deriving from the application of the "reference ratio method".

Pairwise distance PMFs, as used in protein structure prediction, are not physically justified potentials of mean force or free energies

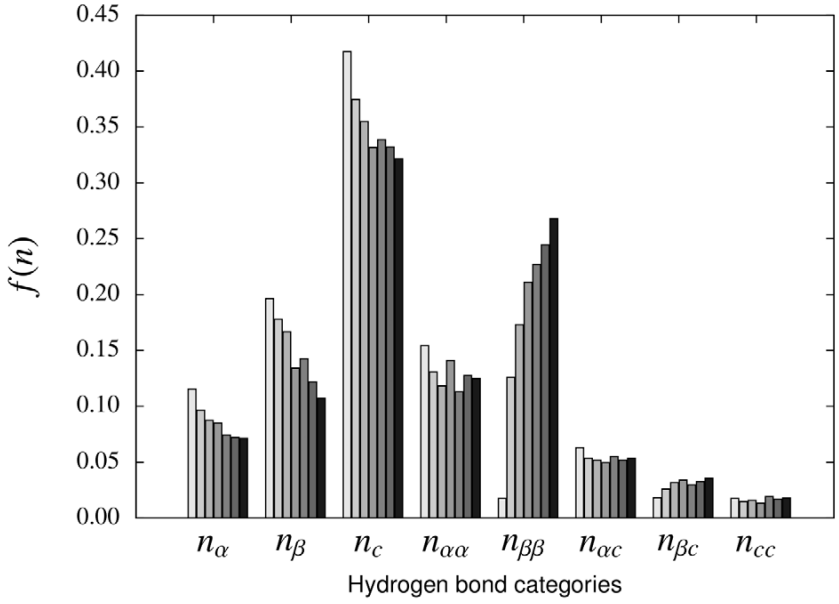


Figure 4. Iterative estimation of a PMF. For each of the eight hydrogen bond categories (see text), the black bar to the right denotes the fraction of occurrence $f(n)$ in the native structure of protein G. The gray bars denote the fractions of the eight categories in samples from each iteration; the first iteration is shown to the left in light gray. In the last iteration (iteration 6; dark gray bars, right) the values are very close to the native values for all eight categories. Note that hydrogen bonds between β-strands are nearly absent in the first iteration (category $n_{\beta\beta}$).
doi:10.1371/journal.pone.0013714.g004

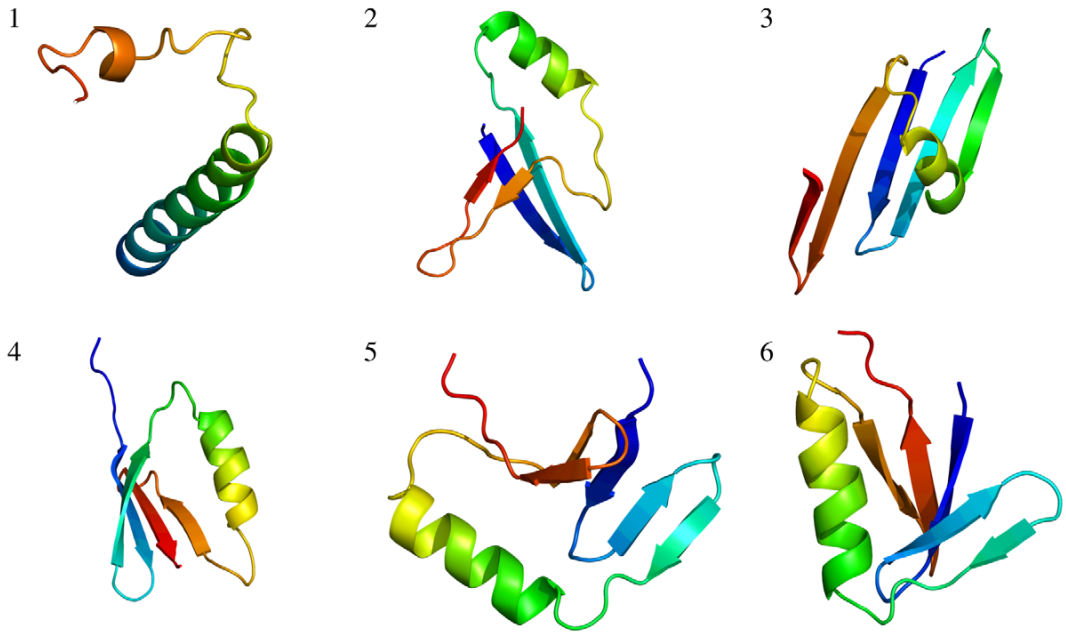


Figure 5. Highest probability structures for each iteration. The structures with highest probability out of 50,000 samples for all six iterations (indicated by a number) are shown as cartoon representations. The N-terminus is shown in blue. The figure was made using PyMOL [64].
doi:10.1371/journal.pone.0013714.g005

[2,13] and the reference state does not depend on external physical considerations; the same is of course true for our generalization. However, these PMFs are approximations of statistically valid and rigorous quantities, and these quantities can be generalized beyond pairwise distances to other aspects of protein structure. The fact that these quantities are not potentials of mean force or free energies is of no consequence for their statistical rigor or practical importance – both of which are considerable. Our results thus vindicate, formalize and generalize Sippl's original and seminal idea [7]. After about twenty years of controversy, PMFs – or rather the statistical quantities that we have introduced in this article – are ready for new challenges.

Materials and Methods

Outline of the problem

We consider a joint probability distribution $Q(X,Y)$ and a probability distribution $P(Y)$ over two variables of interest, X and Y, where Y is a deterministic function of X; we write $Y(X)$ when relevant. Note that because Y is a function of X, it follows that $Q(X)=Q(X,Y(X))$; and if $Y'\neq Y(X)$, then $Q(X,Y')=0$.

We assume that $P(Y)$ is a meaningful and informative distribution for Y. Next, we note that $Q(X,Y)$ implies a matching marginal probability distribution $Q_R(Y)$ (where the subscript R refers to the fact that $Q_R(Y)$ corresponds to the reference state, as we will show below):

$$Q_R(Y)=\int Q(X,Y)\mathrm{d}X$$

We consider the case where $Q_R(Y)$ differs substantially from $P(Y)$; hence, $Q_R(Y)$ can be considered as incorrect. On the other hand, we also assume that the conditional distribution $Q(X|Y)$ is indeed meaningful and informative (see next section). This distribution is given by:

$$Q(X|Y)=\begin{cases}0 & \text{if } Y\neq Y(\mathrm{X})\\ \dfrac{Q(X)}{\int Q(X')\delta(Y(X')-Y)\mathrm{d}X'} & \text{if } Y=Y(\mathrm{X})\end{cases} \quad (10)$$

where $\delta(\cdot)$ is the delta function. The question is now how to combine the two distributions $P(Y)$ and $Q(X)$ – each of which provide useful information on X and Y – in a meaningful way. Before we provide the solution, we illustrate how this problem naturally arises in protein structure prediction.

Application to protein structure

In protein structure prediction, $Q(X,Y)$ is often embodied in a fragment library; in that case, X is a set of atomic coordinates obtained from assembling a set of polypeptide fragments. Of course, $Q(X,Y)$ could also arise from a probabilistic model, a pool of known protein structures, or any other conformational sampling method. The variable Y could, for example, be the radius of gyration, the hydrogen bond network or the set of pairwise distances. If Y is a deterministic function of X, the two variables are called *coarse grained* and *fine grained* variables, respectively. For example, sampling a set of dihedral angles for the protein backbone uniquely defines the hydrogen bond geometry between any of the backbone atoms.

Above, we assumed that $Q(X|Y)$ is a meaningful distribution. This is often a reasonable assumption; fragment libraries, for example, originate from real protein structures, and conditioning on protein-like compactness or hydrogen bonding will thus result in a meaningful distribution. Of course, sampling solely from $Q(X,Y)$ is not an efficient strategy to obtain hydrogen bonded or compact conformations, as they will be exceedingly rare. We now provide the solution of the problem outlined in the previous section, and discuss its relevance to the construction of PMFs.

Solution for a proper joint distribution

A first step on the way to the solution is to note that the product rule of probability theory allows us to write:

$$P(X,Y)=P(Y)P(X|Y)$$

As only $P(Y)$ is given, we need to make a reasonable choice for $P(X|Y)$. We assume, as discussed before, that $Q(X|Y)$ is a meaningful choice, which leads to:

$$P(X,Y)=P(Y)Q(X|Y)$$

In the next step, we apply the product formula of probability theory to the second factor $Q(X|Y)$, and obtain:

$$P(X,Y)=P(Y)\frac{Q(X,Y)}{Q_R(Y)} \tag{11}$$

The distribution $P(X,Y)$ has the correct marginal distribution $P(Y)$.

In the next two sections, we discuss how this straightforward result can be used to great advantage for understanding and generalizing PMFs. First, we show that the joint distribution specified by Eq. 11 can be reduced to a surprisingly simple functional form. Second, we discuss how this result can be used in MCMC sampling. In both cases, expressions that correspond to a PMF arise naturally.

PMFs from combining distributions

Using the product rule of probability theory, Eq. 11 can be written as:

$$P(X,Y)=P(Y)\frac{Q(Y|X)Q(X)}{Q_R(Y)}$$

Because the coarse grained variable Y is a deterministic function of the fine grained variable X, $Q(Y|X)$ is the delta function:

$$P(X,Y)=P(Y)\frac{\delta(Y-Y(X))Q(X)}{Q_R(Y)} \tag{12}$$

Finally, we integrate out the, now redundant, coarse grained variable Y from the expression:

$$\begin{aligned}P(X)&=\int P(X,Y)\mathrm{d}Y\\&=\int P(Y)\frac{\delta(Y-Y(X))Q(X)}{Q_R(Y)}\mathrm{d}Y\\&=\frac{P(Y(X))}{Q_R(Y(X))}Q(X)\end{aligned}$$

and obtain our central result (Eq. 6). Sampling from $P(X)$ will result in the desired marginal probability distribution $P(Y)$. The influence of the fine grained distribution $Q(X,Y)$ is apparent in the fact that $P(X|Y)$ is equal to $Q(X|Y)$. The ratio in this expression corresponds to the usual probabilistic formulation of a PMF; the distribution $Q_R(Y)$ corresponds to the reference state. In the next section, we show that PMFs also naturally arise when $P(Y)$ and $Q(X,Y)$ are used together in Metropolis-Hastings sampling.

PMFs from Metropolis-Hastings sampling

Here, we show that Metropolis-Hastings sampling from the distribution specified by Eq. 11, using $Q(X,Y)$ as a proposal distribution, naturally results in expressions that are equivalent to PMFs. The derivation is also valid if the proposal distribution depends on the previous state, provided $Q(X,Y)$ satisfies the detailed balance condition.

According to the standard Metropolis-Hastings method [60], one can sample from a probability distribution $\Pi(X,Y)$ by generating a Markov chain where each state X',Y' depends only on the previous state X,Y. The new state X',Y' is generated using a proposal distribution $\pi(Y',X'|Y,X)$, which includes $\pi(X',Y'|X,Y)=\pi(X',Y')$ as a special case. According to the Metropolis-Hastings method, the proposal X',Y' is accepted with a probability α:

$$\alpha(X',Y'|X,Y)=\min(1,p),$$

$$p=\frac{\Pi(X',Y')}{\Pi(X,Y)}\times\frac{\pi(X,Y|X',Y')}{\pi(X',Y'|X,Y)} \tag{13}$$

where Y,X is the starting state, and Y',X' is the next proposed state. We assume that the proposal distribution $\pi(X',Y'|X,Y)$ satisfies the detailed balance condition:

$$\pi(X',Y'|X,Y)\pi(X,Y)=\pi(X,Y|X',Y')\pi(X',Y')$$

As a result, we can always write Eq. 13 as:

$$\frac{\Pi(X',Y')}{\Pi(X,Y)}\times\frac{\pi(X,Y)}{\pi(X',Y')}$$

The Metropolis-Hastings expression (Eq. 13), applied to the distribution specified by Eq. 11 and using $Q(X',Y')$ or $Q(X',Y'|X,Y)$ as the proposal distribution, results in:

$$\frac{P(Y')Q_R(Y)Q(X',Y')}{P(Y)Q_R(Y')Q(X,Y)}\times\frac{Q(X,Y)}{Q(X',Y')}$$

which reduces to:

$$\frac{P(Y')}{P(Y)}\times\frac{Q_R(Y)}{Q_R(Y')} \tag{14}$$

Hence, we see that the Metropolis-Hastings method requires the evaluation of ratios of the form $P(Y)/Q_R(Y)$ when $Q(X',Y')$ or $Q(X',Y'|X,Y)$ is used as the proposal distribution; these ratios correspond to the usual probabilistic formulation of a PMF. Finally, when Y is a deterministic function of X, the proposal distribution reduces to $Q(X')$ or $Q(X'|X)$, and Eq. 14 becomes:

$$\frac{P(Y(X'))}{P(Y(X))}\times\frac{Q_R(Y(X))}{Q_R(Y(X'))}$$

Application to radius of gyration and hydrogen bonding

Conformational sampling from a suitable $Q(X|A)$ was done using TorusDBN [45] as implemented in Phaistos [61]; backbone angles (ϕ,ψ and ω) were sampled conditional on the amino acid sequence. We used standard fixed bond lengths and bond angles in

constructing the backbone coordinates from the angles, and represented all side chains (except glycine and alanine) with one dummy atom with a fixed position [61].

For the radius of gyration application, we first determined $Q_R(r_g|A)$ using the multi-canonical MCMC method to find the sampling weights $w(r_g)$ that yield a flat histogram [62]. Sampling from the resulting joint distribution (Eq. 8) was done using the same method. In both cases, we used 50 million iterations; the r_g bin size was 0.08 Å. Sampling from TorusDBN was done conditional on the amino acid sequence A of ubiquitin (76 residues, PDB code 1UBQ).

For the hydrogen bond application, sampling from the PMFs was done in the $1/k$-ensemble [63], using the Metropolis-Hastings algorithm and the generalized multihistogram method for updating the weights [62]. In each iteration i, 50,000 samples (out of 50 million Metropolis-Hastings steps) were generated, and the parameters of the multinomial distribution $Q_{R,i}(H)$ were subsequently obtained using maximum likelihood estimation. Hydrogen bonds were defined as follows: the N,O distance is below 3.5 Å, and the angles formed by O,H,N and C,O,H are both greater than 100°. Each carbonyl group was assumed to be involved in at most one hydrogen bond; in case of multiple hydrogen bond partners, the one with the lowest H,O distance was selected. Each residue was assigned to one of the eight possible hydrogen bond categories ($n_\alpha,n_\beta,n_c,n_{\alpha\alpha},n_{\beta\beta},n_{cc},n_{\alpha c},n_{\beta c}$) based on the presence of hydrogen bonding at its carbonyl group and the secondary structure assignments (for both bond partners) by TorusDBN. The target distribution – the multinomial distribution $P(H)$ used in Eq. 9 – was obtained by maximum likelihood estimation using the number of hydrogen bonds, for all eight categories, in the native structure of protein G (56 residues, PDB code 2GB1). Sampling from TorusDBN was done conditional on the amino acid sequence of protein G; native secondary structure information was *not* used.

Author Contributions

Conceived and designed the experiments: TH JFB. Performed the experiments: MB MP. Analyzed the data: TH MB MP JFB. Contributed reagents/materials/analysis tools: MB MP JP JF CA WB SB. Wrote the paper: TH.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181: 223–230.
2. Moult J (1997) Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 7: 194–199.
3. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15: 2507–2524.
4. Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9: 945–950.
5. Miyazawa S, Jernigan R (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18: 534–552.
6. Miyazawa S, Jernigan R (1999) An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 36: 357–369.
7. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859–883.
8. Chandler D (1987) Introduction to Modern Statistical Mechanics. New York: Oxford University Press, USA.
9. McQuarrie D (2000) Statistical mechanics University Science Books, USA.
10. Finkelstein A, Badretdinov A, Gutin A (1995) Why do protein architectures have Boltzmann-like statistics? Proteins 23: 142–150.
11. Rooman M, Wodak S (1995) Are database-derived potentials valid for scoring both forward and inverted protein folding? Protein Eng 8: 849–858.
12. Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257: 457–469.
13. Ben-Naim A (1997) Statistical potentials extracted from protein structures: Are these meaningful potentials? J Chem Phys 107: 3698–3706.
14. Koppensteiner WA, Sippl MJ (1998) Knowledge-based potentials–back to the roots. Biochemistry Mosc 63: 247–252.
15. Shortle D (2003) Propensities, probabilities, and the Boltzmann hypothesis. Protein Sci 12: 1298–1302.
16. Kirtay C, Mitchell J, Lumley J (2005) Knowledge based potentials: The reverse Boltzmann methodology, virtual screening and molecular weight dependence. QSAR & Combinatorial Sci 24: 527–536.
17. Muegge I (2006) PMF scoring revisited. J Med Chem 49: 5895–5902.
18. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225.
19. Colubri A, Jha A, Shen M, Sali A, Berry R, et al. (2006) Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. J Mol Biol 363: 835–857.
20. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17: 355–362.
21. Eramian D, Shen M, Devos D, Melo F, Sali A, et al. (2006) A composite score for predicting errors in protein structure models. Protein Sci 15: 1653–1666.
22. Rykunov D, Fiser A (2010) New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics 11: 128.
23. Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358: 86–89.
24. Májek P, Elber R (2009) A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. Proteins 76: 822–836.
25. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions1. J Mol Biol 295: 337–356.
26. Gilis D, Rooman M (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence1. J Mol Biol 272: 276–290.
27. Gilis D, Rooman M (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. Protein Eng 13: 849–856.
28. Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci 18: 2550–2558.
29. Chandler D (2005) Interfaces and the driving force of hydrophobic assembly. Nature 437: 640–647.
30. Bowman GR, Pande VS (2009) Simulated tempering yields insight into the low-resolution Rosetta scoring functions. Proteins 74: 777–788.
31. Shmygelska A, Levitt M (2009) Generalized ensemble methods for de novo structure prediction. Proc Natl Acad Sci U S A 106: 1415–1420.
32. Bryngelson J, Wolynes P (1987) Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci U S A 84: 7524–7528.
33. Leopold P, Montal M, Onuchic J (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci U S A 89: 8721–8725.
34. Dill K, Chan H (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4: 10–19.
35. Reith D, Pütz M, Müller-Plathe F (2003) Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24: 1624–1636.
36. Fain B, Levitt M (2003) Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc Natl Acad Sci U S A 100: 10700–10705.
37. Sippl MJ, Ortner M, Jaritz M, Lackner P, Flockner H (1996) Helmholtz free energies of atom pair interactions in proteins. Fold Des 1: 289–98.
38. Zhang C, Liu S, Zhou H, Zhou Y (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 13: 400–411.
39. Cheng J, Pei J, Lai L (2007) A free-rotating and self-avoiding chain model for deriving statistical potentials based on protein structures. Biophys J 92: 3868–3877.
40. Rykunov D, Fiser A (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins 67: 559–568.
41. Bernard B, Samudrala R (2008) A generalized knowledge-based discriminatory function for biomolecular interactions. Proteins 76: 115–128.
42. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275: 895–916.
43. Pearl J (1988) Probabilistic reasoning in intelligent systems. San Francisco: Morgan Kaufmann. pp 108–115.
44. Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10: 139–145.
45. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, et al. (2008) A generative, probabilistic model of local protein structure. Proc Natl Acad Sci U S A 105: 8932–8937.

46. Sippl M, Hendlich M, Lackner P (1992) Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin β4. Protein Sci 1: 625–640.
47. Hamelryck T, Kent J, Krogh A (2006) Sampling realistic protein conformations using local structural bias. PLoS Comput Biol 2: e131.
48. Zhao F, Peng J, Debartolo J, Freed K, Sosnick T, et al. (2010) A probabilistic and continuous model of protein conformational space for template-free modeling. J Comput Biol 17: 783–798.
49. Hamelryck T (2009) Probabilistic models and machine learning in structural bioinformatics. Stat Methods Med Res 18: 505–526.
50. Kullback S, Leibler R (1951) On information and sufficiency. Annals Math Stat 22: 79–86.
51. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.
52. Bowie J, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164–164.
53. Liithy R, Bowie J, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83–85.
54. Buchete NV, Straub JE, Thirumalai D (2004) Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 14: 225–232.
55. Rooman M, Kocher J, Wodak S (1991) Prediction of protein backbone conformation based on seven structure assignments: Influence of local interactions. J Mol Biol 221: 961–979.
56. Kocher J, Rooman M, Wodak S (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 235: 1598–1613.
57. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34: 82–95.
58. Thomas P, Dill K (1996) An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A 93: 11628–11633.
59. Huang S, Zou X (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J Comp Chem 27: 1866–1875.
60. Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice Chapman & Hall/CRC, USA.
61. Borg M, Mardia K, Boomsma W, Frellsen J, Harder T, et al. (2009) A probabilistic approach to protein structure prediction: PHAISTOS in CASP9. In: Gusnanto A, Mardia K, Fallaize C, eds. LASR2009 - Statistical tools for challenges in bioinformatics. Leeds, UK: Leeds University Press. pp 65–70.
62. Ferkinghoff-Borg J (2002) Optimized Monte Carlo analysis for generalized ensembles. Eur Phys J B 29: 481–484.
63. Hesselbo B, Stinchcombe R (1995) Monte Carlo simulation and global optimization without parameters. Phys Rev Lett 74: 2151–2155.
64. Delano WL (2002) The PyMOL Molecular Graphics System. Palo Alto: DeLano Scientific.