PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A simple probabilistic model of multibody interactions in proteins

Kristoffer Enøe Johansson[1] and Thomas Hamelryck[2]*

[1] Section for Biomolecular Sciences, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200, Copenhagen N, Denmark

[2] Section for Computational and RNA biology, Department of Biology, University of Copenhagen, Room 1.2.22, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark

ABSTRACT

Protein structure prediction methods typically use statistical potentials, which rely on statistics derived from a database of know protein structures. In the vast majority of cases, these potentials involve pairwise distances or contacts between amino acids or atoms. Although some potentials beyond pairwise interactions have been described, the formulation of a general multibody potential is seen as intractable due to the perceived limited amount of data. In this article, we show that it is possible to formulate a probabilistic model of higher order interactions in proteins, without arbitrarily limiting the number of contacts. The success of this approach is based on replacing a naive table-based approach with a simple hierarchical model involving suitable probability distributions and conditional independence assumptions. The model captures the joint probability distribution of an amino acid and its neighbors, local structure and solvent exposure. We show that this model can be used to approximate the conditional probability distribution of an amino acid sequence given a structure using a pseudo-likelihood approach. We verify the model by decoy recognition and site-specific amino acid predictions. Our coarse-grained model is compared to state-of-art methods that use full atomic detail. This article illustrates how the use of simple probabilistic models can lead to new opportunities in the treatment of nonlocal interactions in knowledge-based protein structure prediction and design.

Proteins 2013; 81:1340–1350.

Key words: probabilistic models; multibody interactions; Bayesian networks; knowledge-based potentials; protein structure prediction; protein design; visible volume.

INTRODUCTION

The importance of Bayesian probabilistic models is increasing rapidly in structural bioinformatics and structural biology.[1] Examples include statistical superposition of protein structures,[2] inference of protein structure from NMR data,[3,4] and protein design.[5] In protein structure prediction, probabilistic models of local protein structure form an attractive alternative to the use of fragment libraries[6] and side-chain rotamers.[7]

The formulation of probabilistic models of local protein structure can now be considered a solved problem.[6–12] However, this is not the case for nonlocal interactions. Particularly, a nonlocal model is required that can be combined with a local model to form a complete and rigorous model of protein structure.[13]

Probabilistic models of nonlocal interactions are important for the formulation of knowledge-based potentials used in protein structure prediction, such as the potentials of mean force (PMF) and the statistical contact potentials as formulated by Sippl[14,15] and Miyazawa and Jernigan,[16] respectively. Currently, most PMFs consider distance dependent, pairwise atomic contacts.[17,18] Recent important developments in the estimation and formulation of knowledge-based potentials include contrastive divergence[19] and Mullinax and Noid's variational method.[20]

PMFs were originally justified by analogy with the reversible work theorem for liquids, and therefore thought to be limited to pairwise distance-dependent interactions.[15,21,22] However, many PMFs venture beyond the classic formulation, for example, by considering higher order interactions[23–28] or alternative features such as angles.[17,29] Such extensions are in principle perfectly

Additional supporting Information may be found in the online version of this article.

Grant sponsor: Danish Council for Independent Research; Grant number: FTP274-08-0124.

*Correspondence to: Thomas Hamelryck, Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200, Copenhagen N, Denmark. E-mail: thamelry@binf.ku.dk.

Received 13 October 2012; Revised 31 January 2013; Accepted 18 February 2013
Published online 6 March 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24277

valid, as shown by a recent probabilistic formulation of PMFs.[13] This formulation validates and extends these potentials beyond pairwise interactions to arbitrary coarse-grained variables, for example, involving the orientation of amino acid pairs.[29] In addition, the nature of the so-called reference state, necessary for the formulation of a PMF, is now understood.[13]

From a practical point of view, the main goals of computational models are speed and accuracy. To this end, coarse graining of the protein representation continues to be attractive.[30] However, to combine accuracy with coarse graining it is necessary to include higher order interactions.[31] As a result, coarse-grained contact models have been developed that include higher order contacts.[23–28] Like pairwise models, these are formulated as probability tables and in practice seen as limited to three-body or four-body interactions due to lack of data.[32]

Here, we present a probabilistic model of multibody contacts in proteins that does not suffer from an arbitrary upper limit on the number of contacts. The method is formulated as a simple Bayesian network that models the three-dimensional neighborhood of a single amino acid. The model describes the steric and chemical environment using well-known probability distributions. This makes the model completely transparent and applicable as a tool for detailed quantitative analysis of protein structure. The formulation as a Bayesian network enables the application of powerful algorithms for self-consistent parameter estimation and inference. In this article, we present the model, discuss its properties, validate and illustrate its potential, and finally discuss some possible applications.

METHOD AND THEORY

General multibody model

We consider a single amino acid, A_i, at position i in the protein structure and the conditional probability distribution, $P(A_i|\mathbf{A}_{\sim i}, \mathbf{X})$, conditioned on all other amino acids in the sequence, $\mathbf{A}_{\sim i} = \{A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_L\}$, and the structure, $\mathbf{X}$. Here, L is the total number of amino acids in the sequence. This conditional distribution is modeled by a Bayesian network called the *multibody multinomial (MuMu) model* and forms the main topic of this work. We shall discuss it in detail later.

Consider the *joint* conditional distribution, $P(\mathbf{A}|\mathbf{X})$, of an entire sequence, $\mathbf{A} = \{A_1, A_2, \ldots, A_L\}$, consisting of L amino acids. This joint distribution can be factorized into expressions that concern the individual positions, dependent on the remainder of the sequence:[33,34]

$$P(\mathbf{A}|\mathbf{X}) \approx \prod_i P(A_i|\mathbf{A}_{\sim i}, \mathbf{X}) \qquad (1)$$

This factorization is also know as the pseudo-likelihood approximation[35,36] or composite conditional likelihood.[37] Using this approximation, we avoid considering positions as independent while still achieving a factorization that makes the joint distribution tractable. In the following we will discuss, in detail, the residue environment before presenting the probabilistic model.

Residue environment

We consider the first shell of neighbors as observed from the C_β position of the central amino acid. From this observing point, we extract the volume enclosed by the first shell of neighbors together with their numbers and types. In the following section, we will discuss how these features are extracted and modeled as random variables before we discuss the entire Bayesian network.

Visible volume

We apply the visible volume measure of Conte and Smith[38] using the tetrahedral window construction described in the original article and illustrated in Figure 1. This measure reports the volume of empty space as observed from a C_β position and hence the volume available for the side chain atoms attached to the C_β. The volume around the C_β position is divided into four equally sized windows by a tetrahedral construction. The tetrahedron's orientation is determined by the C_α, C_β, C, and N atoms, see Figure 1. As in the original formulation, we neglect the lower window. The three remaining windows, v_0, v_1, and v_2, are modeled as log-transformed volumes by a 3D Gaussian distribution:

$$V = \{\log(v_0), \log(v_1), \log(v_2)\} \sim \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (2)$$

Here, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance matrix. Applying a log-normal model increases the sensitivity for buried (low volume) environments. For Gly, a C_β position is constructed from the backbone atoms.

Neighbor counts

The visible volume construction allows us to define a first shell of neighbors, namely, those that are visible from the C_β position. The first neighbor encountered in a line of sight occludes any more remote neighbors. This construction is similar to the recently published shadow map.[40]

Because we only count the neighbors in the first shell, the model does not rely on the choice of a contact sphere radius or the size of the central amino acid. Conventional burial measures like the coordination number and half-sphere exposure,[41] that count all neighbors in a sphere or half-sphere, are sensitive to the size of the central amino acid.

The neighbors are represented by the chemical groups given in Table I. These groups cluster between two and 11 atoms into chemical relevant entities and enable transfer learning of features that are common to different side chains, for example, the carboxylate group (COO^-)

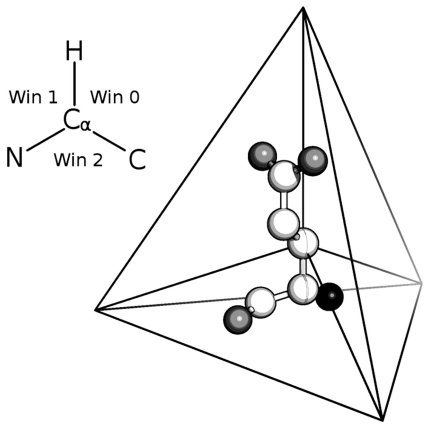

Figure 1

Example of the tetrahedral window construction used in the visible volume calculation. The space around the C_β position of a Glu residue is divided into four equally sized windows by a tetrahedron that is oriented by the C_α, C_β, C, and N atoms. The three upper windows used in this work are schematically illustrated on the left. Figure made in Pymol.[39]

of Asp and Glu. A chemical group is represented by a flat disk perpendicular to the direction of the observing C_β, see Figure 2. In addition to transfer learning, this coarse graining contributes directionality, for example, for the amphiphilic side chain of Thr as illustrated in Figure 2.

The ethyl ($—C_2H_4—$) and propyl ($—C_3H_6—$) groups are merged because they have common chemical properties (sp^3 hybridized carbon chain) but have different disk radii in the contact calculation. The C_α atom is typically contained in these groups (Table I).

The neighbor counts are vectors with 13 nonnegative elements which we model by the multinomial distribution

$$C = \{n_1, n_2, \ldots, n_{13}\} \sim \mathcal{M}(N, p_1, p_2, \ldots, p_{13}) \qquad (3)$$

where $p_1, \ldots, p_{13}$ are the probabilities of the 13 chemical group types and N is the total number of neighbors. Because the multinomial is conditioned on N, we include a separate distribution to model N, see Figure 3.

The backbone conformation, as measured by the ϕ and ψ angles, is an important feature in the evaluation of an amino acid in a given environment. To model these angles, we use the 2D von Mises distribution,[43] an unimodal distribution on the torus, which has previously proven extremely useful to model Φ and ψ angles in continuous space.[6] Backbone information is not included for the terminal positions, which only have a single dihedral angle.

MuMu: A MULTIBODY MULTINOMIAL MODEL

Based on these environment features, we will now present the MuMu model. Following the pseudo-likelihood approximation, Eq. (1), we construct the probability, $P(A_i|\mathbf{A}_{\sim i}, \mathbf{X})$.

The remainder of the sequence and the structure, $\mathbf{A}_{\sim i}, \mathbf{X}$, is represented by the first shell of neighbors, C_i, the total neighbor count, N_i, the visible volume, V_i, and the backbone angles Φ_i and ψ_i:

$$P(A_i|\mathbf{A}_{\sim i}, \mathbf{X}) = P(A_i|C_i, N_i, V_i, \Phi_i, \Psi_i) = P(A_i|E_i) \qquad (4)$$

We will refer to $C_i, N_i, V_i, \Phi_i, \Psi_i$ as the environment, E_i, of residue i.

Table I
Chemical Groups

Group number	Name	Formula	Radius	Amino acids
1	Side chain amide	$—CONH_2$	2.5 Å	Asn, Gln
2	Amine	$—NH_{3+}$	1.8 Å	Lys, N-term
3	Carboxylate	$—COO^-$	2.0 Å	Asp, Glu, C-term
4	Alcohol	—OH	1.3 Å	Ser, Thr, Tyr
5	Thiol and sulfide	—S(H)	1.4 Å	Cys
6	Ethyl	$—C_2H_4—$	1.7 Å	All except Ala, Glu, Gly, Ile, Leu, Gln, Trp
6	Propyl	$—C_3H_6—$	2.5 Å	Glu, Ile, Lys, Leu, Gln, Trp
7	Methyl	$—CH_3$	1.4 Å	Ala, Ile, Leu, Thr, Val
8	Benzene	$—C_6H_{(3\text{-}5)}$	3.0 Å	Phe, Trp, Tyr
9	Imidazole	$—N_2C_3H_3$	2.8 Å	His
10	Guanidine	$—N_3CH_{5+}$	2.5 Å	Arg
11	CSC	$—SC_2H_5$	2.5 Å	Met
12	CN	$—CNH_2—$	1.8 Å	Trp
13	Backbone amide	—CONH—	2.5 Å	All except N-term

Neighbor representation in MuMu. All side chains are represented by one to three chemical groups. The backbone amides are represented by a dedicated group (group 13), while the C_α atoms are included in the side chain ethyl and propyl groups (group 6), with the exception of Ala and Gly. In order not to increase the number of groups unnecessarily, Gly and Ala are only represented by a backbone amide, and a backbone amide combined with a methyl group, respectively. The N and C terminal groups are represented by amine and carboxylate groups, respectively. In the calculations, the groups are represented by a flat disk of the given radius that is perpendicular to the observing direction. The radii are based on Ref. 42. Ethyl and propyl are represented by the same chemical group, but differ in radius.

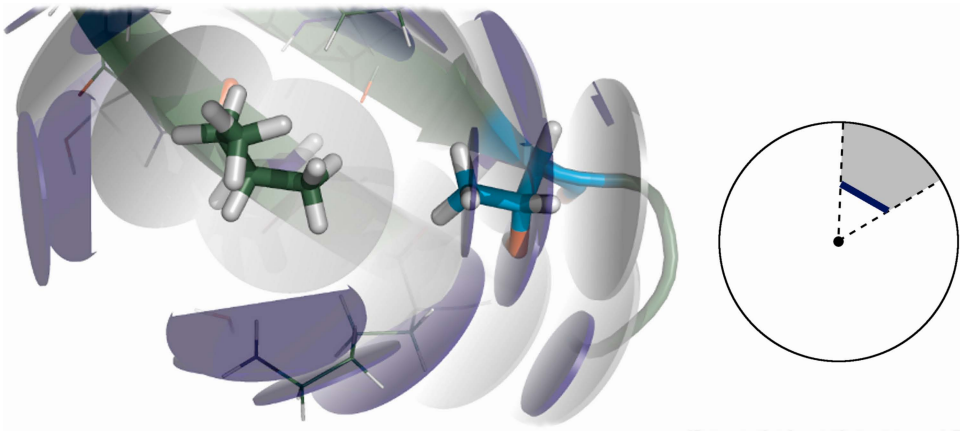

Figure 2

Illustration of the amino acid environment representation in the MuMu model. The chemical group neighbors of an Ile residue (left, green) are represented by shading disks. Backbone amide groups are shown in gray and side-chain groups in blue. The methyl group of the Thr (center, cyan) covers the line of sight to the alcohol group of the same side chain. Thus, the first shell of neighbors contains the methyl group but not the alcohol. The sketch (right) illustrates how a neighboring disk (blue) occludes a fraction of the sphere volume (gray). Figure made in Pymol.[39] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The conditional amino acid distributions, $P(A_i|E_i)$, can be obtained via the product rule

$$P\,(A_i|E_i) \;= \frac{P\,(A_i, E_i)}{P\,(E_i)}. \qquad (5)$$

The joint probability, $P(A_i, E_i)$, can conveniently be formulated as a mixture model[44] by introducing a latent parent variable, H:

$$P(A_i, C_i, N_i, V_i, \Phi_i, \Psi_i) = \sum_{h \in H} P(A_i|h)P(C_i|N_i, h)P(N_i|h)P(V_i|h)P(\Phi_i, \Psi_i|h)P(h) \qquad (6)$$

where the sum runs over all values of H. The size of H determines the number of mixture components which each define a set of parameters for the five distributions: a categorical distribution for the central amino acid, A_i; a multinomial distribution for the neighbor counts, C_i; a categorical distribution for the total number of neighbors, N_i; a 3D Gaussian distribution for the log-transformed visible volume, V_i and a 2D von Mises distribution for the backbone angles, Φ_i, Ψ_i. The MuMu model is shown as a Bayesian network in Figure 3.

The MuMu model is constructed and estimated using the Mocapy++ toolkit for inference of Bayesian networks[45] and public available in Phaistos, a framework for Monte Carlo simulation of protein structure.[46]

Training data

The model is trained on a large data set of globular and soluble protein structures extracted from the PDB. For each residue in a structure, the amino acid type, visible volume, contacts, and backbone conformation are extracted to constitute a data point. We composed a training data set without membrane proteins, disordered proteins, small proteins, or complexes. Furthermore, we exclude structures with ligands and missing atoms because the visible volume measure considers empty space and thus is sensitive to cavities due to missing ligands or atoms. We use structures that are solved using NMR experiments, or crystal structures with a resolution better than 3.8 Å. Results from threading experiments suggest that coarse-grained statistical potentials are fairly insensitive to structure quality,[28] which was also confirmed in our case by tests using a high-quality PISCES data set.[47] Furthermore, the integral nature of volume is

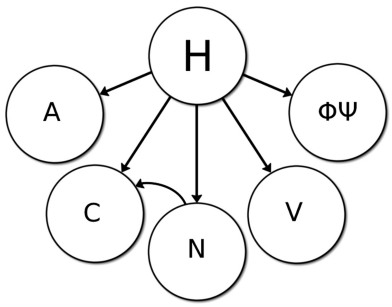

Figure 3

The MuMu model. Circular nodes represent random variables and the arrows encode the conditional independencies. H is a latent variable that all observed variables are conditioned on, A is the central amino acid, C is the contact vector, N is the total number of neighbors, V is the visible volume and Φ, Ψ are the backbone angles. Note that the contact vector, C, is conditioned on the total number of neighbors, N.

particular robust towards positional noise.[38] A non-redundant data set was obtained using BLASTCLUST.[48] This resulted in 276,987 data points of which 5000 random points were reserved for testing purposes. A detailed description of the data set is provided in the Supporting Information together with a complete list of PDB structures (Supporting Information Table SII).

Model estimation

Because we formulate our model as a Bayesian network, we can apply well-established algorithms for self-consistent parameter estimation. We applied the stochastic expectation maximization (EM) algorithm as implemented in the Mocapy++ toolkit.[45] One hundred and fifty models were estimated and the optimal model was selected based on performance and the Bayesian information criteria (BIC).[49] Maximizing the BIC is a well established method for model selection.[50,51] The BIC measure considers the likelihood $P(\mathcal{D}|\Theta_{max})$ of the data, $\mathcal{D}$, given a set of optimized parameters, Θ_{max}, that maximize the likelihood. The BIC measure is calculated as

$$\mathrm{BIC} = \log P(\mathcal{D}|\Theta_{\max}) - \frac{1}{2} M \log(d) \qquad (7)$$

where d is the number of observed environments (training data points) and M is the number of free parameters in the model. This number increases linearly with the size of H, that is, the number of mixture components. Over-fitting the data is avoided due to the second term.[43] Figure 4 shows the BIC measure versus the size of H. The model suggested by the BIC analysis was selected and verified to have the best performance of all models in the amino acid prediction test (Fig. 7). The final model consists of 125 mixture components with 68 parameters in each, resulting in a total of 8500 parameters.

RESULTS AND DISCUSSION

Model analysis

Each of the 125 mixture components describe a characteristic nonlocal amino acid environment by tying together preferences for the central amino acid, visible volume, number of contacts, and backbone dihedral angles (Fig. 3). We will show that these components represent structural motifs which are in many cases well-known. The EM algorithm used for model estimation can be understood as a clustering method that assigns a number of data points to each mixture component and estimates the parameters of the child nodes using these data points. The weight of the component is simply the fraction of data points assigned to the component. For example, the most populated component has 6819 data points assigned to it and correspondingly a weight of $6819/276,987 = 0.025$. This component represents a Val, Ile, or Leu in a β-sheet. It is known that the β-branched amino acids stabilize sheet structure even if solvent exposed[52] and hence, this component captures a known feature of proteins. In the following, we will discuss five illustrative examples of components, for which distributions are shown in Figure 5.

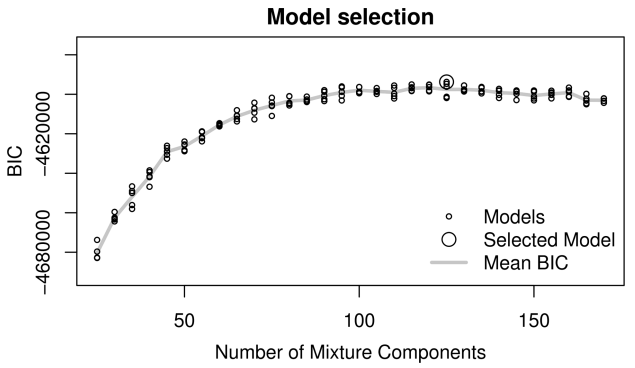

Figure 4

The Bayesian information criterion as a function of model size. The model with maximal BIC is highlighted.

Measuring environment burial

For the sake of discussion, we formally define a buried environment as a mixture component with a low visible volume that is narrowly distributed. We consider the sum of (non-log) window visible volume means, $V_m = \sum_{w=0}^{2} \exp(\mu_w)$, together with the distribution width given as the visible volume differential entropy, S_V.[53] The latter is reported as the more intuitive $\exp(S_V)$. In the MuMu model, this number ranges from 1.1 to 86 where $\exp(S_V) = 1$ represent a lower limit that indicates zero entropy corresponding to zero variance in the visible volume distribution. The component with $\exp(S_V) = 1.1$ is shown in gray in Figure 5 panel **D**. There are 97 components with narrow visible volume distributions ($\exp(S_V) < 20$) and of these, 42 are defined as buried environments with $V_m < 1800$ Å^3.

Confined environment

The purple component represents the most confined environment in the model with the lowest mean volume ($V_m = 574$ Å^3, panel **D**). As expected, the only amino acid that fits into the smallest environment is Gly (panel **A**). The second lowest V_m is equal to 702 Å^3. The corresponding component represents small non-Gly amino acids including Ala, Ser, Thr, Asn, Cys, Val, and Asp (component not shown).

Exposed environment

The gray component has the highest mean visible volume ($V_m = 3475$ Å^3) and represents a completely exposed environment with only three neighboring con-

tacts on average (panel **C**). Accordingly, this environment prefers any of the nonhydrophobic amino acids (panel **A**). This component has a narrow distribution of visible volume (panel **D**) and a backbone preference which indicates an exposed turn motif.

Aromatic rich environment

The green component in Figure 5 has the highest preference for Phe (45%, panel **A**). This corresponds to 941 or 8% of the Phe's in the training data. In general, this component represents large amino acids and accordingly there is a minimum visible volume required. The volume visible through window 0 is higher than that of windows 1 and 2 and hence this component also encodes a preference for the side chain conformation (panel **D** left 2D projection and Fig. 1). By manual inspection of the training data, we confirm that 98% of the data points assigned to this component feature a side chain that occupies window 0.

Disulfide rich environment

The training data contain 1570 observations of Cys participating in a disulfide bridge of which 79% are assigned to four components with different backbone conformations. The red component in Figure 5 is one of these and represents the majority, namely, 40% of the total disulfide observations. Panel **C** shows that observations assigned to this component have on average 15.7 neighbors of which $4.8\% \sim 0.75$ involve a thiol group. Together with panel **A**, this indicates that 70% of the data assigned to this component participate in a disulfide bridge. The visible volume of this component is broadly distributed with $\exp(S_V) = 28$ which, together with the Φ, Ψ distribution, indicates a coil structure.

Pro C-capping motif

The least populated component of the model represents only 62 data points and is shown in yellow in Figure 5. This tiny component is justified in the EM algorithm by covering data points that would otherwise be outliers. This component is a structural analog to the well-known Pro C-capping motif.[54] The central residue of the Pro C-capping motif is positioned at the C-terminal of a solvent exposed α-helix and is followed by a Pro. A characteristic backbone conformation of $\Phi, \Psi = -130^{\circ}, 70^{\circ}$ followed by a Pro gives the motif a characteristic 'S' shape that allows the backbone to return into the structure from the solvent exposed C-terminal of the helix. By inspection of the training data assigned to this component, we confirm that 36 of the 62 data points are followed by a Pro and 26 of these are classified as C-cap by the program DSSP.[55]

When the central amino acid is a His, it can form a hydrogen bond with the backbone oxygen of the residue positioned one helix turn down-stream in the sequence. This requires the side chain to occupy window 1 and hence, this particular window has larger visible volume than the two others (panel **D**, left 2D projection). The narrow distribution of the window 2 volume is also characteristic for helix environments, because this window points toward the helix axis. This demonstrates how the transparency of Bayesian networks can be exploited for automated analysis of protein structures and exploration of the motif vocabulary.

Three of the data points assigned to the yellow component but not classified as Pro C-capping motif (XP) are a variant of the motif, HRP, in which the central amino acid is an Arg that follows a hydrogen bonding and C-capping His and is itself followed by a Pro. In this example, the yellow component can be viewed as a structural generalization of the Pro C-capping motif.

Number of neighbors

The mixture components have on average twelve neighbors. The number of neighbors is in general narrowly distributed with an average variance of 2.6. This variance is significantly smaller than the one expected (12) for the Poisson distribution, which is the most common model for count data. This indicates that the individual mixture components each represent a different but close to fixed number of neighbors. The use of a Bayesian network and a multinomial distribution to model the neighbors makes it possible to consider an arbitrary amount of neighbors. For comparison, a tabulated 12-body method with this representation would require $13^{12} \sim 10^{13}$ parameters, which underlines the power of modeling statistical independencies in the presented way.

Residual probability annotation

Figure 6 visualizes the probability of the native amino acids on a structural representative ensemble of ubiquitin. The ensemble is a collection of X-ray crystal structures that represent the dynamics of ubiquitin.[56] The dark blue residues show that the hydrophobic core has a high probability according to MuMu. Gln-41 is the only amino acid that has a low probability in the structure. A low probability might indicate that the residue takes part in some other aspect of the protein's function such as catalysis or folding. In this case, Gln-41 is highly conserved in the ubiquitin family and is known from NMR experiments to be less ordered than the rest of the protein.[57,58]

Figure 6 also demonstrates that the coarse-grained nature of the method makes it robust towards structural fluctuations. In all members of the ensemble, Gln-41 is recognized as unlikely compared to the other amino acids.

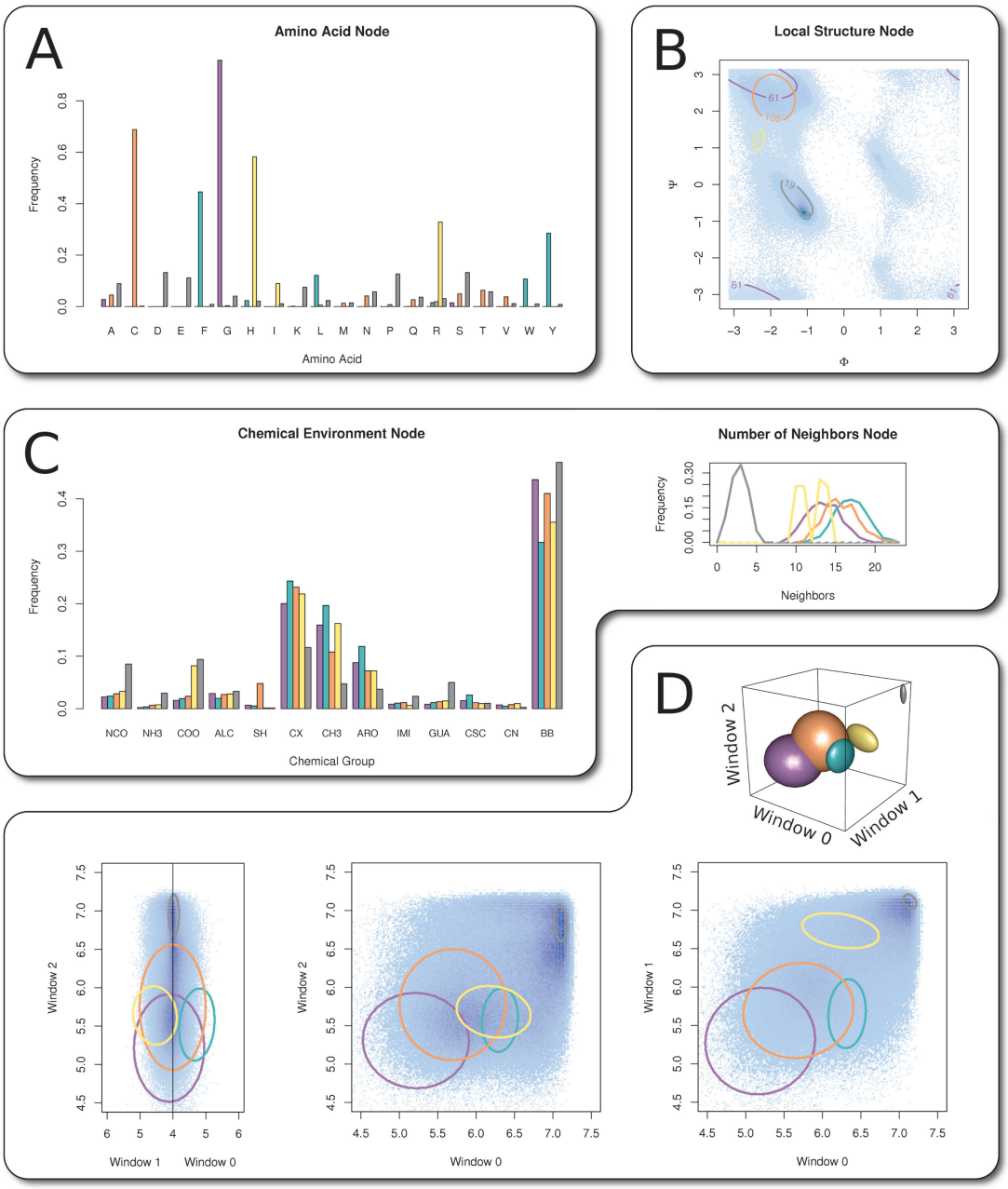

Figure 5

Estimated distributions of five illustrative mixture components of the MuMu model. Components are represented by the colors purple, gray, green, red, and yellow in all panels. Panel **A** shows the categorical distribution of the central amino acid, *A*, panel **B** the 2D von Mises distribution of the local structure, **Φ**, **Ψ**, panel **C** the multinomial distribution of the first shell of neighbors, *C*, together with the categorical distribution of the number of neighbors, *N*, and panel **D** the 3D Gaussian distribution of the log visible volume, *V*. The last includes three 2D projections. Panels B and D are contoured at half of the maximum density and drawn on top of a histogram of the training data plotted in blue. Components are discussed in the text.

VALIDATION

Amino acid prediction

A data point consists of connected observations of amino acid type, *A*, and environment, *E*, which is represented by the three features visible volume, chemical contacts, and backbone conformation. In the first validation of the MuMu model, we attempt to predict the native amino acid type of test data points selected at random and excluded from the training data.

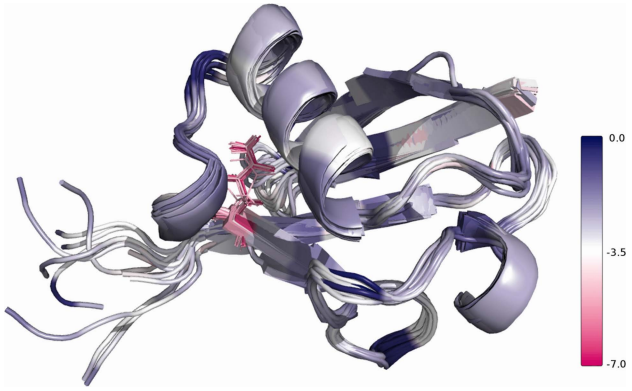

Figure 6

Structural representative ensemble of ubiquitin with residues colored according to the MuMu probability, $\log P(A_i|E_i)$. Log probabilities are indicated in the color bar. Magenta residues have low probability and are thus an unlikely combination of side chain and environment. The side chain of Gln-41 is shown to emphasize that this amino acid has low probability in all structures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In Figure 7, we rank the native amino acid of 5000 test data points. For each test data point, (A_i, E_i), we calculate the probability of the 20 amino acids, $P(A|E_i)$, conditioned on the environment, E_i, and report the rank of the native amino acid, A_i, on the x-axis. If the native amino acid is the most likely in environment E_i, its rank is one. The y-axis gives the fraction of test data points for which the native amino acid is ranked according to the x-value or better. For example, the solid line shows that for 62% of the test data points, the native amino acid is among the five most probable according to MuMu.

Figure 7 shows that $\sim 20\%$ of the test data points are ranked higher than ten which is a poor prediction. Most natural sequences contain amino acids that are suboptimal from a structural point of view[59] and these will be assigned a high rank by the MuMu model. This may explain many of the poor predictions. Gln-41 in Figure 6 is an example of such an structural suboptimal amino acid.

Gly and Pro have distinct preferences for their backbone conformation and, to a lesser extent, visible volume, which make them more easy to predict (dot–dash line in Fig. 7). Polar amino acids are difficult to predict because they often populate the intermediate buried regions which in general have a broader distribution of amino acids. The comparatively poor ranking of Cys might simply be because it is relatively rare, and hence poorly represented in the model.

Comparison with Rosetta

In order to make a performance comparison with the RosettaDesign[60] method, we attempt to predict the native amino acid of all positions of human ubiquitin in Figure 8. As in the previous test, we consider the rank of the native amino acid and report the fraction of correct predictions as a function of rank threshold. The comparison in Figure 8 considers all 76 positions of the PDB structure 1UBQ. The null model simply ranks the 20 amino acids according to the frequency of natural occurrence such that a native Leu is always ranked one.

Rosetta uses full atomic detail of the structure and has in several cases proven a powerful method to predict sequences for a given structure.[59–61] See Supporting Information for details on the Rosetta predictions. As expected, the MuMu model cannot offer quite the same predictive power but it has the benefits of coarse-grained methods, mainly speed. Because of the atomic details modeled by Rosetta, it is necessary to position all side chains at all positions which makes this method $10^3 - 10^5$ times slower than the MuMu model. Considering the three highest ranked amino acids (x-axis in Fig. 8), the null model predicts 22%, the MuMu model 51%, and Rosetta 59% of the native amino acids. As in other studies,[17,62] we expect that the presence of ubiquitin in the large training set is insignificant because this represents only 76 observations out of 276,987 ($< 0.03‰$). Furthermore, the low probability of Gln-41 (Fig. 6) confirms that the model is indeed not overfitted for this particular structure.

Decoy recognition

We compare the MuMu model to the newest versions of the DFIRE[17,63] and DOPE potentials.[18] They are both distance-dependent statistical potentials that use full atomic detail, here representing state-of-the-art within

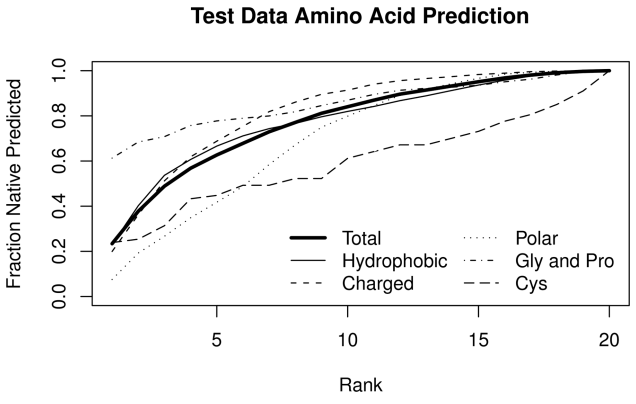

Figure 7

Amino acid prediction of 5000 random test data points. The x-axis shows the rank (lowest is best) of the native amino acid according to MuMu and the y-axis the fraction of data for which the native is predicted as this rank or better. Hydrophobic amino acids are Ala, Phe, Ile, Leu, Met, Val, Trp, and Tyr. Charged amino acids are Asp, Glu, Lys and Arg, and polar are His, Asn, Gln, Ser, and Thr.

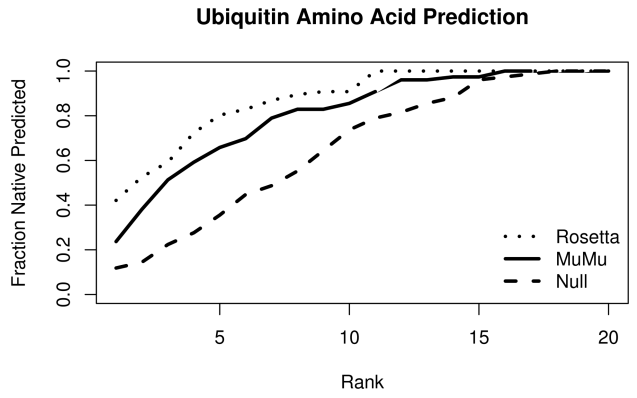

Figure 8

Native amino acid prediction of all positions of human ubiquitin (PDB id: 1UBQ). The MuMu model is compared to Rosetta and a null model (MuMu without environment input). The *x*-axis shows the rank of the native amino acid and the *y*-axis the fraction of data for which the native is predicted as this rank or better.

statistical potentials.[18,64,65] Recognizing a native structure in a large set of computationally generated structures (decoys) is a classic method to validate knowledge-based energies or probabilistic models. This verification method has some limitations[61,66] but is quite informative. We compare with the current state-of-the-art within the field of decoy recognition by applying DFIRE and DOPE potentials to decoy collections generated by iTasser[67] and Rosetta.[68] Decoys generated using other methods are represented by the Gilis standard collection[66] which also includes older Rosetta decoys.

Individual ranking results are given as Supporting Information Table SIV and summarized in Table II. We use the conditional probability of the sequence given the structure, $P(\mathbf{A}|\mathbf{X})$ [Eq. (1)], to assess decoys with the MuMu model. Strictly, this does not evaluate the structures itself but rather the probability of the native sequence given the structure, that is, the likelihood.

Highest emphasis should be put on the Gilis collection because of the large number of decoys.[66] Here, the methods perform similar with DOPE being a bit better (32 native recognized) than the rest. The most challenging decoys in the Gilis standard collection are the Rosetta structures (see Supporting Information Table SIV). Of the 58 native structures in the Rosetta07 collection, only 15 were ranked lowest (most probable) by MuMu compared to 23 for DFIRE2. However, all methods have similar average z-score for the Rosetta07 collection (except for dDFIRE), indicating that the difference is not that pronounced. All methods perform well on the iTasser decoys; MuMu recognizes only a single native less compared to the best methods. In general, it is difficult to prefer one of these methods above the others because their performance varies between the three collections used here. We judge most of the variance in Table II to be insignificant. However, we point out the difference in method complexity as DOPE and DFIRE consider all atoms and contain hundreds of thousands parameters.

Decoys are often generated from high quality target structures which are also preferred for model training. As a consequence, decoy collections usually overlap with training data.[62] In Supporting Information Table SIII, we give the numbers in Table II for the decoy collections without training data homologous. Approximately, one third of the decoys are homologous to a protein in the training set and are hence excluded in Supporting Information Table SIII. Generally, the results are unaltered which indicates that the model does not suffer from overfitting. The training data used with DOPE and DFIRE are not available and hence we could not remove homologues.

Marginal evaluations

To assess the individual node contributions of the MuMu model, we evaluate the decoys only using a subset of the features (panels **B**, **C,** and **D** in Fig. 5) by integrating out all others. For example, the MuMu-contacts (Fig. 5 panel **C**) likelihood is calculated as

$$P(\mathbf{A}|\mathbf{X}) \approx \prod P(A_i|C_i, N_i) \qquad (8)$$

leaving out information on visible volume and backbone angles. Results are given in Supporting Information Table SIII and summarized in Table II. The MuMu-ΦΨ results suggest that Rosetta decoys in general have good local structure. This evaluation ranks in all but one case a Rosetta generated structure lower than the native. In contrast, the iTasser decoys can often be discarded based on the local structure, which shows that Rosetta is better at modeling local structure.

Modeling neighboring contacts by a multinomial distribution as presented in this work is shown to perform

Table II
Decoy Recognition Summary

Decoy collection	Gilis standard	Rosetta07	iTasser refined
Native	41	58	53
Decoys	44381	6960	23298
DOPE	32/−4.1	21/−1.6	45/−3.6
dDFIRE	30/−4.7	12/−0.8	45/−4.7
DFIRE2	28/−3.6	23/−1.7	42/−2.8
MuMu	29/−3.1	15/−1.5	44/−3.4
MuMu-ΦΨ	23/−2.9	1/ 0.7	47/−4.3
MuMu-contacts	15/−2.0	21/−1.7	7/−1.1
MuMu-volume	15/−2.3	18/−1.6	23/−2.3

Summary of the decoy ranking experiment using the DFIRE, DOPE, and MuMu methods. The number of natives ranked highest (correct prediction) is given followed by the average Z-score (lowest is best) of the native structures. The two first rows show the number of decoy set (natives) and total number of decoys in each of the three collections. The three bottom rows are the marginal MuMu evaluations in which all but one observed feature is integrated out.

on par with the much more detailed DOPE and DFIRE methods for the Rosetta07 decoys. MuMu-contacts alone suffice to recognize 21 out of 58 natives in the challenging Rosetta07 collection. The marginal evaluations show that there is little redundancy in our model and that all nodes contribute to the performance of the model.

CONCLUSIONS

We presented a probabilistic model of multibody interactions in proteins based on a Bayesian network. Contrary to common belief, the data available in the current database of protein structures suffices to estimate the parameters of such a model adequately. The presented MuMu model can predict site-specific amino acids from a native environment to a degree that is close in accuracy and orders of magnitude faster than state-of-the-art methods that use full atomic detail. Decoy recognition experiments confirm the quality of the model. Furthermore, the model can be used for detailed quantitative analysis of protein structures, see Figure 5. We conclude with a brief discussion of some possible applications.

The presented formulation enables direct Gibbs sampling of amino acids conditioned on the environment. This can be exploited as a generative proposal distribution in probabilistic protein design based on a Monte Carlo method. Compared to other models in the field of probabilistic protein design,[5,34] the MuMu model is very fast and independent of any rotamer assumptions. In protein design, MuMu could be used prior to side chain construction and detailed evaluation.

PMFs require probabilistic models of nonlocal interactions. To this end, MuMu could be used to formulate a multibody PMF to be combined with a probabilistic model of local structure.[13] As shown in Ref. 13, the formulation of a PMF requires a reference state whose nature is entirely determined by the conformational sampling method adopted. The main challenge in the formulation of a PMF based on MuMu is the development of a second MuMu model concerning the reference state. Such an endeavor is beyond the scope of this article, but quite feasible as shown in Ref. 13.

Finally, MuMu can be used to identify unlikely amino acid environments as illustrated in Figure 6. This could be useful to identify errors in experimental protein structures or to identify unusual residues. In the latter case, an obvious application is the identification of catalytically or functionally important amino acids, as these often occur in unusual, energetically unfavorable environments due to functional requirements.[69]

ACKNOWLEDGMENTS

The authors thank Jesper Ferkinghoff-Borg, Jakob R. Winther, and members of the Hamelryck group for many helpful discussions. Furthermore, we thank Simon Olsson and Jes Frellsen for suggestions on this manuscript.

REFERENCES

1. Hamelryck T, Mardia K, Ferkinghoff-Borg J, editors. Bayesian methods in structural bioinformatics. Statistics for biology and health. New York: Springer; 2012.
2. Theobald DL, Wuttke DS. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian procrustes problem. Proc Natl Acad Sci USA 2006;103:18521–18527.
3. Rieping W, Habeck M, Nilges M. Inferential structure determination. Science 2005;309:303–306.
4. Olsson S, Boomsma W, Frellsen J, Bottaro S, Harder T, Ferkinghoff-Borg J, Hamelryck T. Generative probabilistic models extend the scope of inferential structure determination. J Magn Reson 2011;213:182–186.
5. Fromer M, Yanover C. A computational framework to empower probabilistic protein design. Bioinformatics 2008;24:i214–i222.
6. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. Proc Natl Acad Sci USA 2008;105:8932–8937.
7. Harder T, Boomsma W, Paluszewski M, Frellsen J, Johansson K, Hamelryck T. Beyond rotamers: a generative, probabilistic model of side chains in proteins. BMC Bioinformatics 2010;11:306.
8. Wang Z, Xu J. A conditional random fields method for RNA sequence structure relationship modeling and conformation sampling. Bioinformatics 2011;27:i102–i110.
9. Zhao F, Li S, Sterner BW, Xu J. Discriminative learning for protein conformation sampling. Proteins 2008;73:228–240.
10. Zhao F, Peng J, Debartolo J, Freed KF, Sosnick TR, Xu J. A probabilistic and continuous model of protein conformational space for template-free modeling. J Comput Biol 2010;17:783–798.
11. Lennox KP, Dahl DB, Vannucci M, Tsai JW. Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. J Am Stat Assoc 2009;104:586–596.
12. Lennox KP, Dahl DB, Vannucci M, Day R, Tsai JW. A Dirichlet process mixture of hidden Markov models for protein structure prediction. Ann Appl Stat 2010;4:916–942.
13. Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS ONE 2010;5:e13714.
14. Sippl MJ. 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 1990;213:859–883.
15. Sippl MJ, Ortner M, Jaritz M, Lackner P, Flöckner H. Helmholtz free energies of atom pair interactions in proteins. Fold Des 1996;1:289–298.
16. Miyazawa S, Jernigan RL. Residue—residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623–644.
17. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72:793–803.
18. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
19. Podtelezhnikov AA, Ghahramani Z, Wild DL. Learning about protein hydrogen bonding by minimizing contrastive divergence. Proteins 2007;66:588–599.
20. Mullinax JW, Noid WG. Extended ensemble approach for deriving transferable coarse-grained potentials. J Chem Phys 2009;131:104110.
21. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
22. Koppensteiner WA, Sippl MJ. Knowledge-based potentials—back to the roots. Biochemistry (Mosc) 1998;63:247–252.

23. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci 1997;6:1467–1481.
24. Carter Jr CW, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311:625–638.
25. Mayewski S. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. Proteins 2005;59:152–169.
26. Li X, Liang J. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins 2005;60:46–65.
27. Ngan SC, Inouye MT, Samudrala R. A knowledge-based scoring function based on residue triplets for protein structure prediction. Protein Eng Des Sel 2006;19:187–193.
28. Feng Y, Kloczkowski A, Jernigan RL. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins 2007;68:57–66.
29. Buchete N-V, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. Protein Sci 2004;13:862–874.
30. Tozzini V. Coarse-grained models for proteins. Curr Opin Struct Biol 2005;15:144–150.
31. Liwo A, Ołdziej S, Czaplewski C, Kozłowska U, Scheraga HA. Parametrization of backbone electrostatic and multibody contributions to the UNRES force field for protein structure prediction from ab initio energy surfaces of model systems. J Phys Chem B 2004;108:9421–9438.
32. Lappe M, Bagler G, Filippis I, Stehr H, Duarte JM, Sathyapriya R. Designing evolvable libraries using multi-body potentials. Curr Opin Biotechnol 2009;20:437–446.
33. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
34. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol 2001;306:607–628.
35. Besag J. Statistical analysis of non-lattice data. Statistician 1975;24:179–195.
36. Mardia KV, Kent JT, Hughes G, Taylor CC. Maximum likelihood estimation using composite likelihoods for closed exponential families. Biometrika 2009;96:975–982.
37. Lindsay BG. Composite likelihood methods. Contemp Math 1988;80:221–239.
38. Conte LL, Smith TF. Visible volume: a robust measure for protein structure characterization. J Mol Biol 1997;273:338–348.
39. Schrödinger LLC. The PyMOL molecular graphics system. Version 1. 2r2, 2009.
40. Noel JK, Whitford PC, Onuchic JN. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. J Phys Chem B 2012;116:8692–8702.
41. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 2005;59:38–48.
42. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. Structure 1994;2:641–649.
43. Mardia KV, Taylor CC, Subramaniam GK. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. Biometrics 2007;63:505–512.
44. Bishop C. Pattern recognition and machine learning. Information science and statistics. Heidelberg: Springer; 2006.
45. Paluszewski M, Hamelryck T. Mocapy++—a toolkit for inference and learning in dynamic Bayesian networks. BMC Bioinformatics 2010;11:126.
46. Boomsma W, Frellsen J, Harder T, Bottaro S, Johansson KE, Tian P, Stovgaard K, Andreetta C, Olsson S, Valentin J, Antonov LD, Christensen AS, Borg M, Jensen JH, Lindorff-Larsen K, Ferkinghoff-Borg J, Hamelryck T. J Comput Chem 2013, DOI: 10.1002/jcc.23292.
47. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
49. Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–464.
50. Chickering DM, Heckerman D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Mach Learn 1997;29:181–212.
51. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 1998;41:578–588.
52. Minor DL, Kim PS. Measurement of the β-sheet-forming propensities of amino acids. Nature 1994;367:660–663.
53. Ahmed NA, Gokhale DV. Entropy expressions and their estimators for multivariate distributions. IEEE Trans Inf Theory 1989;35:688–692.
54. Prieto J, Serrano L. C-capping and helix stability: the pro C-capping motif. J Mol Biol 1997;274:276–288.
55. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
56. Lange OF, Lakomek NA, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 2008;320:1471–1475.
57. Briggs MS, Roder H. Early hydrogen-bonding events in the folding reaction of ubiquitin. Proc Natl Acad Sci USA 1992;89:2017–2021.
58. Lakomek NA, Farès C, Becker S, Carlomagno T, Meiler J, Griesinger C. Side-chain orientation and hydrogen-bonding imprint supramotion on the protein backbone of ubiquitin. Angew Chem Int Ed 2005;44:7776–7778.
59. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. Principles for designing ideal protein structures. Nature 2012;491:222–227.
60. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
61. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332:449–460.
62. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
63. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008;17:1212–1219.
64. Fasnacht M, Zhu J, Honig B. Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci 2007;16:1557–1568.
65. Gniewek P, Leelananda SP, Kolinski A, Jernigan RL, Kloczkowski A. Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. Proteins 2011;79:1923–1929.
66. Gilis D. Protein decoy sets for evaluating energy functions. J Biomol Struct Dyn 2004;21:725–736.
67. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS ONE 2010;5:e15386.
68. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmström L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with rosetta@home. Proteins 2007;69:118–128.
69. Beadle BM, Shoichet BK. Structural bases of stability function tradeoffs in enzymes. J Mol Biol 2002;321:285–296.