Published in final edited form as:

Proteins. 2013 May ; 81(5): 841–851. doi:10.1002/prot.24242.

Extracting knowledge from protein structure geometry

Peter Røgen* and

Department of Mathematics, Technical University of Denmark, Building 303B, DK-2800 Kongens Lyngby, Denmark

Patrice Koehl

Department of Computer Science and Genome Center, University of California, Davis, CA 95616

Peter Røgen: Peter.Roegen@mat.dtu.dk; Patrice Koehl: koehl@cs.ucdavis.edu

Abstract

Protein structure prediction techniques proceed in two steps, namely the generation of many structural models for the protein of interest, followed by an evaluation of all these models to identify those that are native-like. In theory, the second step is easy, as native structures correspond to minima of their free energy surfaces. It is well known however that the situation is more complicated as the current force fields used for molecular simulations fail to recognize native states from misfolded structures. In an attempt to solve this problem we follow an alternate approach and derive a new potential from geometric knowledge extracted from native and misfolded conformers of protein structures. This new potential, MPP, has two main features that are key to its success. Firstly, it is composite in that it includes local and non local geometric information on proteins. At the short range level it captures and quantifies the mapping between the sequences and structures of short (7-mer) fragments of protein backbones through the introduction of a new local energy term. The local energy term is then augmented with a non local residue-based pairwise potential, and a solvent potential. Secondly, it is optimized to yield a maximized correlation between the energy of a structural model and its RMS to the native structure of the corresponding protein. We have shown that MPP yields high correlation values between RMS and energy and that it is able to retrieve the native structure of a protein from a set of high-resolution decoys.

Keywords

Pair-potential; Protein configuration space; Solvent exposure; Protein structure refinement; Knowledge-based potential; Comparative modeling; Protein structure prediction

Introduction

Proteins are the essential tools inside the cells that perform nearly all cellular functions. Just like macroscopic tools, it is their shapes that define their functions. Molecular biology has recently undertaken the vast challenge of finding the structures of all known proteins, in hope to unravel this relationship between geometry and function and learn in the process how cells work. Determining experimentally the structure of a protein at the atomic level however is not an easy task: This can be indirectly deduced from the fact that we currently know millions of protein sequences but less than hundred thousand protein structures. In parallel, prediction of the structure of a protein from first principles is not easier: direct applications of the ideas that have been used for modeling small chemical systems have not

*Corresponding author; Peter.Roegen@mat.dtu.dk, Phone: +45 45253044.

yet been successful on these much larger molecules. Recent reports on the advancement of ab initio techniques clearly show that the protein structure prediction community is making progress, but that the quality of the models they generate do not meet yet the stringent accuracy requirements to become useful to the biologists¹. If modeling and experiments cannot solve the structure of all proteins on their own, the hope today is that the solution will come from a combination of these two approaches: this has led to the development of prediction techniques such as homology modeling that rely heavily on the knowledge contained in the database of known (experimental) protein structures. In this paper, we focus on the extraction of such knowledge: more specifically, we show how we can extract information from the geometry of a protein structure to derive a (knowledge-based) potential that is able to identify near native conformations for a protein with high accuracy and that is as smooth as possible.

Nearly all structure prediction techniques proceed in two steps: generation of many models for the protein of interest, followed by an evaluation of all these models to identify those that are native-like. In theory, the second step is easy, as native structures correspond to minima of their free energy surfaces. In practice however, the situation is more complicated. Proteins are too complex to be treated exactly using quantum mechanics models and the current force fields used for simulations fail to recognize native states from misfolded structures. The most successful approaches to this evaluation problem are therefore knowledge based.

The main idea behind knowledge-based potentials is that native protein structures exhibit non-random geometric characteristics that can be used as signatures for correct structures if they can be quantified correctly. Many measures of the geometry of a protein structure have been taken into account in that context, such as the local geometry of dihedral angles, the accessibility to the solvent, packing, the geometry of residue pairs or atom pairs, side-chain orientation preferences,...(see for example the recent reviews of Skolnick² and Rykunov and Fiser³ and references therein). The idea of converting these measures into a score originates from the work of Tanaka and Scheraga⁴; it has been subsequently formalized in many different ways, including using the quasi-chemical approximation^{5,6}, the physical concept of potential of mean force in relation to Boltzmann's law⁷, or through pure statistical considerations⁸⁻¹⁰; the scores are then referred to as "potential of mean force" (PMF), or "statistical potential", depending on the underlying model. Most of these approaches attempt to quantify the non-random properties of native protein structures by comparing them with a reference state (see the pioneering work of Sippl⁷ in this field). The validity of the approximations behind these methods and more importantly the legitimacy of their applications to large bio-molecular systems are regularly discussed in the literature (see for example¹⁰⁻¹⁵). Despite all the doubts that have been raised, the resulting knowledge-based potentials are widely used in the computational biology community for multiple applications, including structure prediction¹⁶, identification of errors³, and assessment of protein-ligand¹⁷ and protein-protein interactions¹⁸.

An alternate approach to using a reference state is to explicitly design a knowledge-based potential for optimal recognition of native structures versus misfolded conformations. Methods that implement this approach optimize a scoring function such that it provides a significant difference between the score of the native structure and the scores of misfolded conformations (also called decoys) for the same structure, using a database of native protein structures and a corresponding database of misfolded conformations as training sets. They mainly differ in their definitions of a meaningful difference between native structures and decoys. The energy gap is probably the simplest form of such a difference. If we define the scores of the native structure and of a decoy conformation as E_N and E_D , respectively, a

significant energy gap can be obtained by optimizing a Z-score $\mathcal{Z}$ defined as $\mathcal{Z} = \frac{E_N - \langle E_D \rangle}{\sigma(E_D)}$ where $\langle E_D \rangle$ and $\sigma(E_D)$ are the average and standard deviation of the energy values for the decoys considered, respectively¹⁹. Alternatively, it is possible to enforce that E_N be smaller than all values E_D using linear programming (see for example Rajgaria, McAllister, and Floudas²⁰). Note that these techniques single out a unique native structure and would not recognize near-native conformations. To alleviate this problem, Qiu and Elber²¹ expanded the concept of native structure by letting a set of native-like structures all serve the role of native conformations in the inequalities that impose higher energy values for decoys.

Ideally, a knowledge-based potential should assign scores to a protein structure model such that the higher the score, the less native-like the model is, where native-like is measured in terms of root mean square (RMS) deviation of the model from the native structure. If this condition is satisfied then the potential is expected to detect near native conformations even when the native conformation is not present; in addition, such an ideal potential could then be used for model refinement. In practice however, there is no guarantee that such a relationship exists. Knowledge-based potentials are usually derived from experimental, native protein structures only and as such do not include information on non native states. Many potentials have a karst like energy surface, with many high energy values and only a few low energy values (among which the native state is found), and very little correlation between energy and RMS², especially for high RMS values (i.e. greater than 4 Å). Potentials that optimize Z-scores or impose that the energy of the native structure be the lowest do not differentiate between decoys.

There have been many attempts to design improved statistical potentials that would generate a meaningful ranking between decoys. The most obvious choice in that regard is to design a potential or score that reflects the distance between the decoy and the native structure. The protein structure prediction package Rosetta²² for example includes a comprehensive statistical potential energy function that combines sequence dependent and sequence independent terms²³. The weights of these terms have been optimized on a training set by fitting the total energy of a decoy to a function $g(RMS)$ where RMS is the root mean square distance between the decoy and the corresponding native structures. The function g is defined as $g(x) = 4$ if $x < 4$, $g(x) = 8$ if $x > 8$, and $g(x) = x$ otherwise. The two cutoffs at 4 Å and 8 Å are set to mean that decoys within 4 Å of the native structure are equally good, while decoys distant by more than 8 Å are all equally bad. We note that this potential is relevant mostly for ab initio protein structure prediction problems as it is not expected to do better than 4 Å resolution. Similar ideas were incorporated into the structure prediction program TASSER²⁴. Its energy function TOUCHSTONE II includes ten different terms, some statistical (pairwise potentials), some based on physical interactions (such as electrostatics)²⁵. Similarly to Rosetta, the weights of these terms have been optimized to maximize the correlation between the energy values for the decoys and their RMS distance to the native structure (using a function similar to $g(RMS)$ presented above, with the upper threshold moved from 8 to 10 Å) and to maximize an energy gap between the native state and an ensemble of unfolded states²⁵. In other variants that follow the same ideas, such as QMEAN²⁶, the distance between a decoy and the native structure is not computed from the RMS but with the global distance test score GDT_TS²⁷. The corresponding optimized potentials have proved useful for refinement of near native models such as those derived from homology modeling²⁴ or from NMR data²⁸; to our knowledge however, these potentials have not been consistently successful for either refining or detecting models that differ from the native structure by more than 2 Å.

In this study, we present a novel knowledge based potential whose energy landscape is sculpted to give good correlation values between energy and RMS in the whole range from 1

to 4 Å. This potential, which we name MPP for Metric Protein Potential, combines a novel local twice continuously differentiable term that compares the geometry of each overlapping 7-mer in the model under consideration with the configuration space accessible to 7-mer fragments with similar sequences in a set of known protein structures, a twice continuously differentiable pairwise potential for non local interactions, and a continuous and piecewise smooth potential to account for solvent effects, that is based both on accessible surface area and occupied volume. The parameters of all terms in MPP are re-fined according to geometry: we optimize the shape of the potentially local minima around the native structures of a set of proteins by maximizing the correlations between energy and RMS. Optimization and testing are both performed on the Titan high resolution decoy dataset ²⁰.

Materials and Methods

Local geometry: quantifying the conformations of contiguous 7-mer fragments

Local protein structure prediction usually does not exceed 85 % accuracy. This is a direct consequence of the degeneracy of the mapping between local sequences and local structures for proteins: it is well known for example that identical pentapeptides have been found to adopt different conformations in proteins ²⁹. There has been a lot of interest however to study this mapping ³⁰. Sequences with ambivalent structures for example have been characterized as conformational switches of proteins ³¹, while sequences with consistent structures are used as seeds for protein structure prediction programs such as ROSETTA ³². Our work is an extension of these studies: we define a local potential that quantifies the mapping between local sequences and local structures, for fragments of size 7 residues. Any such 7-mer M is characterized by its conformation C and its sequence S .

We characterize the geometry of C as follows. First, we use a local smoothing of its backbone configuration using a procedure introduced by Larsen and Røgen ³³ as it is known to provide a good separation of secondary structure types ³⁴. Briefly, given the positions of the carbon alpha atoms, c_i^α , $i = 1, \dots, 7$ in C , we calculate

$$p_i = \frac{c_{i-2}^\alpha + bc_{i-1}^\alpha + ac_i^\alpha + bc_{i+1}^\alpha + c_{i+2}^\alpha}{a + 2b + 2}$$

for $i = 3, 4, 5$, where $a = 2.1$ and $b = 2.4$ have been chosen to minimize the total chain curvature ³³ in different proteins. The fragment $p_3p_4p_5$ is then characterized by its length $L = \|p_5 - p_3\| + \|p_4 - p_3\|$ and its distance excess $DE = \|p_5 - p_3\| - L$ which essentially measures its curvature. The conformation C is then defined as one point in the geometric $L - DE$ -plane, which we re-scale to treat L and DE equally.

We characterize the sequence S as follows. It is well known that many proteins with dissimilar sequences are found to share a common fold, as evidenced in the Protein Data Bank. This has led to the concept of reduced amino acid alphabets, wherein similar amino acids are clustered together. It is usually not clear however to what extent this reduction should be applied. To circumvent this ambiguity, we consider a set of 12 alphabets A_k of different cardinality k , with $k \in 4, \dots, 14, 20$, based on groupings defined by Solis and Rackovsky ³⁵. The sequence S is then represented by 12 different sequences S_k , based on its translations in the corresponding alphabets A_k .

The local energy mapping between sequences and structures of 7-mer fragments is defined as follows. We extract all n observed conformations C_i , $i = 1, \dots, n$, of a S_k in a database of known protein structures. Note that n usually increases as k decreases for any sequence S , as

lower k means more degeneracy in sequence space. If n is zero, the energy map for S_k is set to be zero. In all other cases, we project the conformations C_i on the $L - DE$ plane which has been discretized with a fine uniform Cartesian grid of 80×80 points. Each conformation C_i is

projected on its nearby grid points at distance r using the smooth kernel $1 - 3\frac{r^2}{R^2} + 2\frac{r^3}{R^3}$ for $r < R = 0.1$ and 0 (zero) otherwise, with the corresponding smoothed values added to the current count values V on the grid. Once all C_i have been projected on the grid, the set of grid points r_{ij} with non zero values $V(r_{ij})$ forms a subset $\Omega(S_k)$. The grid value $V(r_{ij})$ for a point r_{ij} inside $\Omega(S_k)$ is normalized to minus its fraction of the total grid count. This distribution is converted to an energy value using a Boltzmann-like relation, i.e. by taking minus the log of the probability. This leads to problems however when the probability is equal to 0 (i.e. for positions outside of $\Omega(S_k)$). One solution is to add pseudo-counts⁷. We choose instead a linear approximation of the log rule, as it provides smoother gradients for the final potential. Let d_{ij} be the shortest distance in the $L - DE$ plane between each grid point r_{ij} and all the data points associated with S_k . We set

$$W(S_k, r_{ij}) = \frac{(d_{ij}/R - 1/2)^2}{d_{ij}/R + 1/2}$$

for $d_{ij} < R/2 = 0.05$ and 0 otherwise. We call $W^{-1}(0)$ the configuration space of the sequence S_k .

The full energy map $E(S_k)$ is a smoothed approximation to the grid values $V(r_{ij}) + W(r_{ij})$ using a cubic spline with uniform knot vector. By construction $E(S_k)$ is a statistical potential on the configuration space of the 7-mer S_k , with a correction outside this space where it is set to be the shortest distance to known data points. We note that E is twice continuously differentiable. The fine Cartesian grid ensures that if $E(S_k)$ is built from one data point then this point is a global minimum and a line search following the gradient direction from arbitrarily chosen starting points will end very close to the data point.

The total local energy of a protein P_N with N residues is computed as the sum of the $N - 6$ contributions of its 7-mers, with each 7-mer j with conformation C_j contributing 12 values based on the 12 translations $S_{j,k}$ of its sequence S_j in the different amino acid k -alphabets:

$$E_{local}(P_N) = \sum_{j=1}^{N-6} \sum_{k=4, \dots, 14, 20} w_k^{local} E(S_{j,k}, r(C_j))$$

where $r(C_j)$ is the position of C_j in the $L - DE$ plane and the 12 weights $w_k^{local} \geq 0$ express the relative importance of the various levels of sequence similarity considered; they are determined by the final sculpting optimization (see below).

Non local geometry: a pairwise potential

We encode the non-local geometry of the protein structure into a twice continuously differentiable residue-based, distance-dependent, knowledge-based potential computed from the positions of the C_α atoms. This potential is not based on a statistical sampling of known structures; instead, it is defined as follows. For a given pair of amino acid type i, j , the non local energy function E_{NL} is set to be a cubic uniform b-spline with 6 degrees of freedom called the knots of the spline. These knots control the shape of the smoothed potential which is set to level out between 0 and $4.5 \text{ \AA}$ (where there are usually no observation), to be

unrestricted between 4.5 Å and 10 Å and to reach zero smoothly for $r = 10$ Å. The 210 different pairwise potential functions $E_{NL}^{i,j}(r)$ are then fully defined by $210 * 6$ values. These values are considered as parameters that are derived in the sculpting process (see below). We note that the third derivative of each b-spline is a piecewise constant function and linear in the knots. Therefore the sum of the squared discontinuities of the third derivatives of the splines is a (fixed) quadratic form in the knots. This will be used to regularize the final optimization problem such that both the constant zero solution is unique if data are missing and wild fluctuations are penalized.

Finally, to ensure that this potential models long range interactions only, it is multiplied by a switching function f that is based on the distance excess along the smoothened backbone for the protein considered. For two residues at position i and j in the sequence of the protein, whose smoothed C_α are located at positions $\mathbf{p}_i$ and $\mathbf{p}_j$ respectively, we compute both the end

to end distance $d = \|\mathbf{p}_i - \mathbf{p}_j\|$ and the length $L = \sum_{k=i}^{j-1} \|\mathbf{p}_k - \mathbf{p}_{k+1}\|$. Defining $r = L - d$, the switching function is defined as:

$$s(r) = \begin{cases} 0 & r < 1.5 \\ 6(1.5-r)^3(r^2 - 5.5r + \frac{23}{3}) & 1.5 \leq r \leq 2.5 \\ 1 & r > 2.5. \end{cases}$$

The characteristic values 1.5 and 2.5 for this switching function have been chosen such that residue pairs in the same secondary structure element are ignored and a pair on each side of a half turn is counted with full weight.

Modeling solvent effects

The local and non-local energy terms defined above capture the geometry of the protein that reflects its internal energy. Protein stability however is strongly influenced by its environment, i.e. mostly by the solvent. A major role of the solvent is defined through the so-called hydrophobic effects, that sees hydrophobic residues decreasing their contacts with water. To quantify the hydrophobic effect, Lee and Richards introduced the solvent-accessible surface ³⁶. Their ideas were later refined by Eisenberg and McLachlan ³⁷, who introduced the concept of a solvation free energy for large biomolecules, computed as a weighted sum of the accessible areas of all their atoms i . It is not clear, however, if surface area is the correct geometric measure to use in this context: there is some evidence that for small solutes at least, the hydrophobic term is not proportional to the surface area, rather to the solvent excluded volume of the molecule ³⁸. Current models for the non-polar part of the solvent energy include both a surface-based term and a volume-based term ³⁹. We have therefore developed a simplified solvation energy function that depends both on the surface area and volume of each residue in the protein. For a protein with N residues, it is defined as:

$$E_{Solvation} = \sum_{i=1}^N (w_{aa(i)}^{area} SA_i + w_{aa(i)}^{volume} Vol_i)$$

where SA and Vol are the accessible surface areas and solvent-excluded volumes of the residues of the protein when it is represented as a union of N balls of radius 5 Å, centered at the location of the C_α atoms, and $aa(i)$ is the amino acid type of residue i . The surface areas and volumes are computed analytically using the program AlphaVol ⁴⁰.

The solvation energy defined above is not a statistical potential: it is parametric however and depends on the 40 weights w^{area} and w^{volume} that are optimized as described below.

Sculpting the complete potential MPP for linear Energy-RMS relation

The final energy $MPP = E_{Local} + E_{NL} + E_{Solvation}$ is a linear function of its 12 (local) + 210 * 6 (non local) + 40 (solvation) = 1312 parameters. We attempt to sculpt the energy surface to become a linear cone around each native structure using RMS as the measure of distance to the native structure. Ideally, for any protein P whose native structure (that may not be known) is C_{nat} and for which we have a set of model structures C_j , we would like this potential to satisfy:

$$MPP(C_j) - MPP(C_{nat}) = \alpha RMS(C_j, C_{nat})$$

with $\alpha \approx 1$ but fold dependent. Starting from a database of M proteins P for which both the native structure and a large collection of near-native models (often referred to as decoys) are known, we search for the 1312 energy parameters and the $M \alpha_m$'s that minimize the sum of squared errors on these equations and thus solve a least squares problem, with the additional linear constraints that $0 \leq w_k^{local}$ for $k = 4, \dots, 14, 20, \frac{1}{4} < \alpha_m < 4$ for all $m = 1, \dots, M$, $\sum_m \alpha_m = M$, and that the 210 pair potentials are non-negative at zero distance ($0 \leq E_{NL}^{i,j}(0)$ for all amino acid pairs (i, j)). Note that if all the constructed linear Energy-RMS relations have slope one our objective function maximizes the average Energy-RMS correlation; in practice however the correlation values and parameters α s are fold dependent.

We further regularize the objective function by adding a quadratic form expressing the sum of the squared discontinuities of the spline functions representing the pair potentials.

Data sets, training sets and test sets

We distinguish three categories of protein sets: the data sets, that are used to generate the potentials, the training sets, that are used to sculpt the potentials to maximize the correlation between energy and RMS, and the test sets, that are used to test their performances.

We have chosen version 2.4 of the CATH database⁴¹ as our data set. While more recent versions of CATH are readily available, we have used the release 2.4 so as to artificially limit the coverage of known protein structures in the training set (see below). We have included all 19,273 gap-free domains present in CATH2.4 on which the backbone smoothing procedure can be performed (failure arises when they are missing residues in the chain, or when the procedure leads to self-intersection). From these domains we have extracted 2,785,978 7-mers. To ensure that these 7-mers have good geometry, we impose two additional constraints. Firstly, we require that the distance between two consecutive C_α be in the range 3.75Å to 3.86Å; 93% of the distances in the selected 7-mers satisfy this criteria. Note that by applying this filtering, we lose all peptide bonds that are in the Cis configuration. Secondly, we bin the $L - DE$ plane into 200×200 bins and throw away the 1% of the data that lie in the lowest populated bins. After these additional filters, our set of known 7-mers contains 1,831,958 conformations representing 420,441 distinct 7-mer sequences (in the 20-letter alphabet). In native structures the range of values for L is 2.5 Å to 7 Å and the range of values for DE is -0.55 Å to 0 Å. We re-scale this rectangle to a unit square to treat the two measures equally. It should be noted that we did not filter the CATH domains to only consider a non-redundant set (i.e such that no two domains share a high sequence identity). A high density of data in a given region may therefore be an actual representation of the geometry of proteins, or simply the result of a bias in the data set. We did not try to correct for such possible biases.

We have chosen the Titan high resolution decoy (TitanHRD) database as our training set. TitanHRD contains 1400 decoy collections, based on 1400 non-homologous proteins, each with between 500 and 1600 decoy structures²⁰. For 1326 of these decoy collections we were able to select 400 decoys at random and run successfully our procedure for smoothing the backbone (where again success means that we could generate a smoothed chain without self intersection). For most of these decoys the RMS with respect to the native structure is in the range 1 to 4 Å.

To test the performance of our new potential MPP we divided the 1326 curated decoy collections described above into two groups: a training set of 1176 decoy collections and a test set of 150 decoy collections. The test set was chosen to have the same distribution of known 7-mers as the training set.

Results

Sculpting MPP: importance of regularization

We build our combined potential MPP in two stages. Firstly, the knowledge-based elements of the local potential are computed from a subset of the CATH 2.4 database⁴¹ described above. In the second stage, all parameters of the local, non local and solvent energy terms together with the fold dependent slopes of the linear Energy-RMS relations are refined simultaneously to minimize the squared deviations between the total energy values MPP and the desired conical energy surfaces on a large training set. We have used a subset of the TitanHRD database²⁰ for that purpose.

The local part of the potential is the sum of the contributions of all (overlapping) 7-mers in the structure. As this potential is knowledge-based, it may be incomplete and is set to zero if the sequence of a fragment is not “known”. The performance of MPP is therefore expected to depend on the fraction of known 7-mers, FKS, in the test proteins and this dependence is most significant in the 20 amino acid alphabet. We note that the proteins included in our training set and those included in our test set cover the whole spectrum of possible values for FKS, with approximately 30 % of these proteins having an FKS of 0, i.e. none of their fragments are present in the CATH2.4 data set, while less than 5 % have an FKS of 1.0, with at least one copy of all their fragments in CATH2.4.

In addition to the linear fitting, the sculpting procedure also includes a regularization that penalizes the sum of the squares of the discontinuities of the uniform b-spline functions representing the non-local, pair potentials. We have observed that different weights given to the regularization process lead to changes in the performance of MPP, as illustrated in figure 1.

The regularization of the pair-potential improves the Energy-RMS correlations especially for folds with low correlation values. For the optimal value of regularization, the average Energy-RMS correlation is high; 0.82 on the training set and 0.79 on the test set, matching with the average value of 0.79 reported by Floudas and colleagues²⁰.

Local and non local knowledge in protein structures

Figures 2 and 3 show examples of the resulting local and non-local components of the potential, respectively.

The local potential described above maps sequences with structures for small 7-mer fragments. It is designed to penalize any situation in which the structure of a 7-mer lies outside of the configurational space associated with its sequence, as observed in the CATH2.4 database. Figure 2 shows the example of a fragment that has primarily been seen

as short and straight in three main regions that we identify as corresponding to an alpha helix and two coiled configurations. Note, that the top left alpha helix region is most populated but contributes energetically very little compared to the regions outside the configuration space. When the 7-mer geometry falls outside the unit square we extend the definition of the local energy function and chooses it to be the sum of the energy of the nearest boundary point plus the distance to this boundary point. The extended local energy under performs slightly on the 26% of the decoys where it is active.

The non local potential is a residue-based, distance-dependent potential. It is not however a statistical potential; instead, it is sculpted to detect non-native conformations, where non-native is measured in terms of RMS. Interestingly, the resulting potential $E^{i,j}$ captures the expected behavior for the residue types i and j . For example, we show in figure 3 that two tryptophane residues prefer to be apart, with their C_α distant by at least 5.5 Å while aspartic acid and arginine, two charged residues with opposite charges, have a preference for close contacts, with a peak at around 4.5 Å.

Building up MPP: importance of each individual energy term

We repeated the optimization procedure described above for each individual term included in MPP independently, as well as for different combinations of these terms. Results are shown in figure 4.

With no surprise, the performance of the local term in our energy function is strongly dependent on FKS. The value of the Energy-RMS correlation increases first sharply as FKS increases from 0 to 0.3 and slowly levels off, reaching an average value of 0.86 when FKS is 1.0. Both the non-local energy and the solvation energy have little to no dependence on the fraction of known 7-mers, FKS.

MPP, which combines the local, non-local, and solvation potential, is found to outperform each of its components. For FKS above 0.2, the use of MPP on decoys leads to average correlations between energy and RMS in the range 0.8 to 0.9.

It should be noted that figure 4 illustrates the performance of MPP on the same dataset that was used to optimize the potential; as such, it does not fully reflect its possible applications to new proteins. To assess the latter, we tested MPP on an independent test set of 150 folds, also derived from the TitanHRD database but independent from the 1176 folds of the training sets. Results are shown on figure 5.

As expected, the performances on the test set are lower than on the training set. The differences depend on FKS. For low (<0.2) and high (>0.7) FKS values, representing 42% and 48% of the folds respectively, MPP performs as well on the two sets. For intermediate values of FKS however, MPP does not perform as well on the 150 folds of the test set. While we cannot exclude that this is just a reflection that the decoy sets included in the test set are more difficult to analyze, it is reasonable to assume that our procedure has over-learned the training set.

Finally, we assessed the performances of the different components of MPP after it has been optimized and regularized. Results on the test set and training set are given in table 1; these results are to be compared with those obtained with each component being optimized independently, and presented in figure 4. We note first that none of the three individual terms (local, non local, and solvation) dominates the others: they all have similar relative performances. In parallel, all three terms contribute to the performance of the full potential MPP: removing any of the three potentials lead to lower correlation values and more false positives (i.e. decoys whose energy values are lower than the energy of the native structure).

The local energy term based on 7-mers performs almost identically to when it is optimized independently, as reported in figure 4. Interestingly, we find that the 4 and 20 amino acid alphabets dominate the local potential: the optimal parameters of the local energy term are high and almost equal for these two alphabets, and almost zero for the other alphabets. In addition, the relative weights of the k-alphabets are basically unchanged when the local energy term is optimized on its own, or optimized as part of MPP. In the 4-letter alphabet³⁵, the amino acids are gathered into four groups, namely Gly and Pro, hydrophobic (Cys, Phe, Ile, Leu, Met, Val, Trp and Tyr), mostly polar (Asp, Asn, Ser and Thr) and mostly charged (Ala, Glu, His, Lys, Gln and Arg). The amino acids in these four groups have, with a few exceptions, low, high, low, and low propensities for forming extended structures and low, high, low, and high propensities for forming helices, respectively. The dominance of this alphabet is most likely related to the importance of the partitioning of amino acids according to their hydrophobic and hydrophilic natures. In addition, the FKS values are larger for a reduced alphabet. It is unclear however why the other alphabets with between 4 and 20 letters have much lower weights. We are currently investigating this point.

In comparison the performance of the solvation potential drops slightly when it is not optimized alone. The drop in performance for the non-local term is the largest, from correlation coefficients in the range 0.7–0.75 down to the range 0.5–0.58.

Detecting the native structure in a set of decoy structures

A significant difference between the least square procedure we use for optimizing our potential and the procedure developed by Floudas and colleagues²⁰ is that we have not penalized individual decoys for having lower energy values than the native structure. It is therefore a priori not clear whether our optimized potential is well suited for native structure recognition. We checked this on the 150 folds included in our test set. Results are shown in figure 6.

In line with the results on Energy-RMS correlation, the local potential by itself is already performant and identifies the native structures in most cases. When the FKS drops below 15 %, the local potential provides very little information and the other energy terms (non-local and solvation) palliate to some extent this deficiency. Again, the complete potential MPP outperforms each of its components. Overall, we find that the native structure has the lowest energy in 73% of the 150 cases included in the test set.

We have shown above that the regularization of the spline functions that form the nonlocal part of MPP improves the correlations between energy and RMS values (see figure 1). We find that it is also true for smaller values of the regularization constant when MPP is tested with respect to its ability to detect the native structure, as illustrated in figure 7. Interestingly, the regularization narrows the gap between the performances on the training and test sets.

With a medium regularization, MPP assigns the lowest energy to the native structure in 75.3% of the test cases, which is exactly the same value that was reported by Rajgaria, McAllister, and Floudas²⁰. The reasons for the differences in the importance of regularization for setting a high Energy-RMS correlation or for native structure detection are unclear.

An independent test of MPP

We have tested our energy functions on the decoy sets of Park & Levitt⁴² as found on the Decoy-R-Us web repository⁴³. These decoy sets differ significantly from the Titan decoys. While the latter contain decoys that are usually close to the native structures (i.e. with RMS below 5 Å), the former include test sets that have been generated to cover more broadly the

configuration spaces accessible to the proteins. We considered six sets from the Park and Levitt decoy sets, each with 600–700 decoys that cover structures with RMS in the range 0 (native) to 10 Å from the native structure. They were generated using a 4-state off lattice model. All decoy structures have the native secondary structures.

We sculpted MPP on the Titan training set and tested it on these six decoy sets. Results are given in table 2. The MPP values always identify the X-ray structure as the structure with the lowest energy; interestingly, the local energy by itself is as performant as the complete potential in identifying the native structure. The Energy-RMS correlations are similar, and sometimes higher than those found with popular statistical potentials, such as PAT/ME⁴⁴, RAPDF⁹, DFIRE⁴⁵, with the statistical potential developed by Lu and Skolnick⁴⁶, as well as with the standard comprehensive potential included in ROSETTA3.4⁴⁷. These correlation values however are significantly lower than the ones we obtained with the Titan test sets (see figure 5).

Discussion

The knowledge-based potential MPP presented in this study has two important features. Firstly, it captures and quantifies the mapping between the sequences and structures of short (7-mer) fragments of protein backbones through the introduction of a new local energy term. Secondly, it is optimized to yield a maximized correlation between the energy of a structural model and its RMS to the native structure of the corresponding protein. We have shown that MPP yields high energy - RMS correlations and is able to retrieve the native structure of a protein from a set of decoys, both on a test set that is similar to the training set (from the TitanHRD database), and on an independent test set (the 4-state reduced decoy set of Park and Levitt).

The local potential is the most significant element of our potential. It is a combination of a knowledge-based term that quantifies the configurational space of a 7-mer fragment in a protein, and a geometric term when a given fragment falls outside this configuration space. The second term represents to our knowledge a new measure of the structural flexibility that is permitted; it is found to dominate the contribution of our local potential. It should be noted that the size and measures of the configurational space are knowledge-based, and therefore dependent on the quality of the data set used to derive the potential. We have characterized this effect by measuring the efficiency of our potential as a function of the “known” fragments, i.e. fragments that have an equivalent in the data set. As expected, we found the our local potential performs poorly when only a small fraction of the 7-mer fragments in a protein are known. We are currently investigating ways to improve this, by defining for examples fragment “synonyms”, when the actual fragment is not present.

Our non-local residue-based pairwise potential differs from all other similar potentials in that it is not statistical. Instead, each pair potential is modeled as a uniform spline function with six degrees of freedom, whose values are derived by optimization (the “sculpting” procedure). The optimization process leads to potentials that are consistent with the nature of the amino acids involved. We are currently working on extending this idea to other non-local potentials, such as an all-atom potential.

The local and nonlocal terms of our potential are two times continuously differentiable spline functions. It is consequently easy to compute their gradients as well as a full Hessian matrix for the complete potential MPP. These gradients can then be combined with efficient energy minimization algorithms and be used for protein structure refinement. We are currently investigating this option.

Acknowledgments

We acknowledge support from the NIH. We thank Britt Park, Michael Levitt, Ram Samudrala and Christodoulos Floudas for making their decoy sets available for testing.

References

1. Zhang Y. Protein structure prediction: When is it useful? *Curr Opin Struct Biol.* 2009; 19:145–155. [PubMed: 19327982]
2. Skolnick J. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol.* 2006; 16:166–171. [PubMed: 16524716]
3. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics.* 2010; 11:128–138. [PubMed: 20226048]
4. Tanaka S, Scheraga H. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules.* 1976; 9:945–950. [PubMed: 1004017]
5. Miyazawa S, Jernigan R. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 1985; 18:534–552.
6. 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. [PubMed: 8604144]
7. Sippl M. 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. [PubMed: 2359125]
8. Simons K, 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. [PubMed: 9149153]
9. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol.* 1998; 275:895–914. [PubMed: 9480776]
10. 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. [PubMed: 21103041]
11. Rooman M, Wodak S. Are database-derived potentials valid for scoring both forward and inverted protein folding? *Prot Eng.* 1995; 8:849–858.
12. Thomas P, Dill K. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol.* 1996; 257:457–469. [PubMed: 8609636]
13. Ben-Naim A. Statistical potentials extracted from protein structures: are these meaningful potentials? *J Chem Phys.* 1997; 107:3698–3706.
14. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct? *Protein Science.* 1997; 6:676–688. [PubMed: 9070450]
15. Furuichi E, Koehl P. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins: Struct Func Genet.* 1998; 31:139–149.
16. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 2006; 15:2507–2524. [PubMed: 17075131]
17. Fan H, Schneidman-Duhovny D, Irwin J, Dong G, Soichet B, Sali A. Statistical potential for modeling and ranking of protein-ligand interactions. *J Chem Inf Model.* 2011; 51:3078–3092. [PubMed: 22014038]
18. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. *Biophys J.* 2003; 84:1895–1901. [PubMed: 12609891]
19. Mirny L, Shakhnovich E. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol.* 1996; 264:1164–1179. [PubMed: 9000638]
20. Rajgaria R, McAllister S, Floudas C. A novel high resolution Ca-Ca distance dependent force field based on a high quality decoy set. *Proteins: Struct Func Genet.* 2006; 65:726–741.

21. Qiu J, Elber R. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins: Struct Func Genet.* 2005; 61:44–55.
22. Rohl CA, Strauss CE, Misura KMS, Baker D. Protein structure prediction using Rosetta. *Methods Enzymology.* 2004; 383:66–93.
23. Simons K, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct Func Genet.* 1999; 34:82–95.
24. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci (USA).* 2004; 101:7594–7599. [PubMed: 15126668]
25. Zhang Y, Kolinski A, Skolnick J. Touchstone II: A new approach to ab initio protein structure prediction. *Biophys J.* 2003; 85:1145–1164. [PubMed: 12885659]
26. Benkert P, Tosatto S, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Struct Func Bioinfo.* 2008; 71:261–277.
27. Zemla A, Venclovas C, Moulton J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins: Struct Func Genet.* 1999; S3:22–29.
28. Lee S, Zhang Y, Skolnick J. Tasser-based refinement of NMR structures. *Proteins: Struct Func Genet.* 2006; 63:451–456.
29. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci (USA).* 1984; 81:1075–1078. [PubMed: 6422466]
30. Bystroff C, Simons K, Han K, Baker D. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol.* 1996; 7:417–421. [PubMed: 8768900]
31. Young M, Kirshenbaum K, Dill K, Highsmith S. Predicting conformational switches in proteins. *Protein Science.* 1999; 8:1752–1764. [PubMed: 10493576]
32. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol.* 1998; 281:565–577. [PubMed: 9698570]
33. Lindorff-Larsen K, Røgen P, Paci E, Vendruscolo M, Dobson C. Protein folding and the organization of the protein topology universe. *Trends in Biochem Sci.* 2005; 30:13–19. [PubMed: 15653321]
34. Røgen P, Karlsson P. Quantifying the relative positions of proteins secondary structure elements. *Geometrica Dedicata.* 2008; 134:91–107.
35. Solis A, Rackovsky S. Optimized representations and maximal information in proteins. *Proteins: Struct Func Genet.* 2000; 38:149–164.
36. Lee B, Richards FM. Interpretation of protein structures: estimation of static accessibility. *J Mol Biol.* 1971; 55:379–400. [PubMed: 5551392]
37. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature (London).* 1986; 319:199–203. [PubMed: 3945310]
38. Lum K, Chandler D, Weeks JD. Hydrophobicity at small and large length scales. *J Phys Chem B.* 1999; 103:4570–4577.
39. Wagoner J, Baker N. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc Natl Acad Sci (USA).* 2006; 103:8331–8336. [PubMed: 16709675]
40. Edelsbrunner H, Koehl P. The geometry of biomolecular solvation. *Discrete and Computational Geometry (MSRI Publications).* 2005; 52:243–275.
41. Orengo C, Michie A, Jones D, Swindells M, Thornton J. CATH: A hierarchic classification of protein domain structures. *Structure.* 1997; 5:1093–1108. [PubMed: 9309224]
42. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol.* 1996; 258:367–392. [PubMed: 8627632]
43. Samudrala R, Levitt M. Decoys 'R' Us: A database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.* 2000; 9:1399–1401. [PubMed: 10933507]
44. Summa C, Levitt M, DeGrado W. An atomic environment potential for use in protein structure prediction. *J Mol Biol.* 2005; 352:986–1001. [PubMed: 16126228]

45. 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. [PubMed: 12381853]
46. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Struct Func Genet.* 2001; 44:223–232.
47. Leaver-Fay A, MMT, Lewis S, Lange O, Thompson J, Jacak R, KKK, Renfrew P, Smith C, Sheffler W, Davis I, Cooper S, Treuille A, Mandell D, Richter F, Ban Y, Fleishman S, Corn J, Kim D, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek J, Karanicolas J, Das R, Meiler J, Kortemme T, Gray J, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 2011; 487:545–74. [PubMed: 21187238]

Figure 1. Effect of regularization on the performance of MPP

MPP, which includes a local, non-local, and a solvation energy term, is optimized on a training set, under varying regularizations of the discontinuities of the non-local b-spline functions. The corresponding average Energy-RMS correlations computed over the training set (dashed line, red), the test set (dashed-dotted line, green), the combined training+test set (solid line, black), and the set of folds with no known 7-mers (dotted line, blue) are plotted against the weight given to the regularization.

Figure 2.
An example of the level curves of a 7-mer local potential.

Figure 3.

Three examples of the non-local pair-potential spline functions after optimization: Trp-Trp as a dashed line in red, Ala-Lys as a solid line, blue, and Asp-Arg as a dashed-dotted line, in green.

Figure 4. Different components of MPP are optimized and tested on the 1176 decoys of the training sets; the corresponding average Energy-RMS correlation values are plotted as functions of the fractions of known 7-mers (FKS). All curves have been smoothed with a sliding window of 0.1 FKS. Note that the performance of the local potential drops to 0.08 when no 7-mers are known.

Figure 5. The performance of MPP

The average energy-RMS correlation values computed over the training set (solid line, black) and the test set (dashed line, red) are plotted as functions of the fractions of known seven-mers (FKS). Both functions are smoothed with a sliding window of size 0.06.

Figure 6. Detecting native structures in a bag of decoys

The fractions of native structures whose energies are lower than the energies of all their corresponding decoys are plotted against the fraction of known 7-mers (FKS). The different curves correspond to different components of MPP; they have been smoothed with a sliding window of 0.06, to improve readability.

Figure 7.
The fractions of native structures in the training set (solid line, black) and the test set (dashed line, red) whose energies are lower than the energies of all their corresponding decoys are plotted against the weight given to the regularization in the optimization of MPP.

Table 1

Performance of the different components of the regularized energy function MPP

Component ^c	Correlation ^a			Rank (%) ^b		
	All	FKS<0.25	FKS 0.25	All	FKS<0.25	FKS 0.25
Local	0.50 (0.52) ^d	0.12 (0.11)	0.78 (0.83)	18 (18)	40 (40)	2.1 (0.49)
Non Local	0.52 (0.57)	0.51 (0.57)	0.53 (0.58)	10.9 (7.6)	10.0 (8.4)	11.6 (7.0)
Solvation	0.61(0.67)	0.61 (0.65)	0.61 (0.67)	8.2 (7.1)	9.7 (8.1)	7.1 (6.4)
Local+Non Local	0.68 (0.72)	0.50 (0.56)	0.81 (0.85)	6.2 (4.3)	12.5 (9.5)	1.7 (0.25)
Local+Solvation	0.73 (0.76)	0.61 (0.63)	0.83 (0.87)	4.8 (4.4)	10.7 (9.9)	0.57 (0.22)
Non Local + Solvation	0.71 (0.75)	0.67 (0.74)	0.71 (0.76)	4.1 (2.4)	4.7 (2.8)	3.4 (2.1)
All	0.79 (0.82)	0.71 (0.75)	0.84 (0.88)	2.4 (1.4)	5.1 (3.2)	0.53 (0.14)

^a Average correlation coefficients between energy and RMS. Results are reported for all decoys (All) and then broken down into those with a fraction of known seven mers below 0.25 (FKS<0.25), and the others (FKS>0.25).

^b Average ranking, defined as the percentage of decoy structures with lower energy values than the native structure.

^c All combinations of the three main components of the MPP potential are considered.

^d Values are given for the test set and the training set (in parenthesis).

Table 2

Energy-RMS correlations on the 4 state reduced decoy set

PDB ID	FKS	MPP	Local energy	PAT/ME ^{a,e}	RAPDF ^{b,e}	Lu ^c	DFIRE ^{d,e}	Rosetta3 ^f
1ctf	0.26	0.55	0.55	0.72	0.73	0.6	0.7	0.64
1r69	0.96	0.66	0.69	0.66	0.7	0.5	0.65	0.71
1sn3	0.81	0.67	0.72	0.41	0.47	0.5	0.34	0.52
2cro	0.97	0.67	0.68	0.65	0.76	0.7	0.75	0.66
4pti	0.96	0.58	0.65	0.34	0.47	0.5	0.46	0.47
4rxn	0.58	0.73	0.74	0.46	0.58	0.6	0.66	0.57

^aPAT/ME is a statistical potential that captures the preference for atoms to be in a given environment, defined by the number of atoms in that environment and their chemical nature ⁴⁴.

^bRAPDF is an all-atom statistical pairwise potential ⁹.

^cLu and Skolnick ⁴⁶ derived an all-atom statistical potential similar to RAPDF. It differs however in that it is optimized to yield large Z-scores for native structures compared to ensembles of decoy structures. Results are derived from table III of Lu and Skolnick ⁴⁶.

^dDFIRE is another all-atom statistical pairwise potential; it differs from RAPDF in its choice of a reference state ⁴⁵.

^eThe results reported here for PAT/ME, RAPDF, and DFIRE are derived from the original paper describing PAT/ME ⁴⁴.

^fRosetta3 scores were computed using the program “score” from the ROSETTA3.4 suite of programs ⁴⁷, using the standard potential and the standard weights “score12”.