PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities

El-Ad David Amir,[†] Nir Kalisman,[†] and Chen Keasar*

Department of Computer Science, Ben-Gurion University of the Negev, Israel

ABSTRACT

Rotatable torsion angles are the major degrees of freedom in proteins. Adjacent angles are highly correlated and energy terms that rely on these correlations are intensively used in molecular modeling. However, the utility of torsion based terms is not yet fully exploited. Many of these terms do not capture the full scale of the correlations. Other terms, which rely on lookup tables, cannot be used in the context of force-driven algorithms because they are not fully differentiable. This study aims to extend the usability of torsion terms by presenting a set of high-dimensional and fully-differentiable energy terms that are derived from high-resolution structures. The set includes terms that describe backbone conformational probabilities and propensities, side-chain rotamer probabilities, and an elaborate term that couples all the torsion angles within the same residue. The terms are constructed by cubic spline interpolation with periodic boundary conditions that enable full differentiability and high computational efficiency. We show that the spline implementation does not compromise the accuracy of the original database statistics. We further show that the side-chain relevant terms are compatible with established rotamer probabilities. Despite their very local characteristics, the new terms are often able to identify native and native-like structures within decoy sets. Finally, force-based minimization of NMR structures with the new terms improves their torsion angle statistics with minor structural distortion (0.5 Å RMSD on average). The new terms are freely available in the MESHI molecular modeling package. The spline coefficients are also available as a documented MATLAB file.

Proteins 2008; 72:62–73.

Key words: Ramachandran plot; rotamer; optimization; refinement; assessment; CASP; MESHI.

INTRODUCTION

Rotatable torsion angles (TAs) are the major degrees of freedom that enable protein flexibility. Thus, they have attracted intensive research over the last four decades. Since the pioneering work of Ramachandran and his coworkers, a series of computational and statistical studies have shown that adjacent TAs are highly coupled.[1–5] Certain combinations, such as the right-handed helix or valine's most favorable rotamer, are very common. Other combinations are absent from protein structures. Further, the trends observed in large databases with thousands of proteins and hundreds of thousands of residues are mirrored in each of the known protein structures.

Modern X-ray crystallography uses energy functions and other restraints that are based on TAs, but the last steps of refinement usually remove these restraints.[6,7] Thus, statistics of TA values observed in high resolution structures can be used in the analysis of newly determined structures. On one hand, residues with "forbidden" TA combinations (e.g. at the lower right region of the Ramachandran plot) suggest a structure of low quality.[6,8–12] On the other hand, residues with "unfavorable" TAs were shown to be deserving special attention as they are likely to be involved in the protein's function.[11,13]

Unlike crystallography, protein NMR spectroscopy experiments do not allow direct TA determination. The TAs in NMR structures result from an optimization process that takes into account both the NMR constraints and an energy function. Knowledge extracted from high quality X-ray structures may be used to estimate the quality of the NMR structures,[14–16] or to be directly embedded in the energy function.[17–19] Biasing either the energy function or the search procedure towards preferred TA values is also a common practice in protein structure prediction.[20–22]

Quite a few TA energy terms have been proposed so far. They differ in their fundamental assumptions (mainly physics vs. knowledge-based), the specific TAs they handle, and their implementation details. Physics-based force fields typically use one-dimensional sum-of-cosines

Grant sponsors: The GeneFun consortium of the European Commission (LSHG-CT-2004-503567) and Israeli Science Foundation (289/06).

[†]Authors El-Ad David Amir and Nir Kalisman contributed equally to this work.

*Correspondence to: Dr. Chen Keasar, Department of Computer Science, Ben-Gurion University of the Negev, POB 653, Be'er Sheva, Israel. E-mail: chen.keaser@gmail.com

Received 22 May 2007; Revised 10 September 2007; Accepted 2 October 2007
Published online 10 January 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21896

functions (see Ref. 23 and references therein) though $\{\varphi,\psi\}$-cross terms are starting to emerge.[24] Their apparent difficulty in reproducing the observed TA distributions resulted in a flourish of knowledge-based terms that represent either the probability[7,21,25,26] or the propensity[27,28] of residues to adopt a given TA combination. At least one leading group actually uses both types of knowledge-based terms.[22]

Most knowledge-based terms operate on at least two TAs from the same residue as warranted by the strong coupling between TAs in protein residues. The coupling of TAs from successive residues was also studied.[12,28] Side-chain TAs and their coupling to the backbone conformations are often handled by rotamer libraries.[29,30]

The implementation details of energy terms have two major implications: evaluation speed and differentiability, which in turn determine which optimization schemes can be used with them. Lookup table implementations provide rapid energy calculations but only discontinuous first derivatives.[7,18,22] These are clearly the implementations of choice in force-free schemes such as most Monte-Carlo search algorithms.[31] The discontinuous forces, however, may lead to poor convergence in force-driven schemes.[17] The multi-dimensional sum of Gaussian wells provides the opposite compromise, perfect differentiability at the expense of evaluation speed.[17] Recently, Fujitsuka *et al.*[26] presented an implementation that is both efficient and continuously differentiable. They used a bi-cubic spline interpolation of the lookup table for a probability-based φ/ψ term. The spline interpolation ensures continuity of the forces. Their treatment of the side-chains, however, is discrete.

In this study, we present a generic approach for fully differentiable TA energy terms that account for the coupling of all the TAs in a residue. This approach is based on a modular set of multi-dimensional TA energy terms that represent both probability and propensity. Full differentiability is ensured by the use of multi-dimensional cubic splines with periodic boundary conditions. We demonstrate the accuracy, utility, and some limitations of the new energy terms in a diverse series of tests that include comparison with a higher resolution energy term, comparison with rotamer probabilities known from the literature, discrimination of native and native-like structures among diverse decoy sets, and force-based minimization of NMR-derived structures. We argue that the inclusion of these terms may be beneficial for various applications in protein structure prediction, model refinement, and molecular dynamics.

METHODS

Torsion angles database

Our database includes 177,443 residues from a non-redundant set of 850 high quality protein structures.[30] We excluded from the database incomplete residues and residues that include atoms with a high (>40) temperature factor. Residues at chain termini and residues that flank disordered regions are likewise excluded, as they do not have a full set of φ and ψ TAs.

Distances in torsion space

Let $\vec{x} = (x_1,\ldots,x_\mathrm{n})$ and $\vec{y} = (y_1,\ldots,y_n)$ be vectors of TAs. Their distance is defined as:

$$D(\vec{x},\vec{y}) = \sqrt{\sum_{1}^{n}\{\min(|x_i - y_i|, 2\cdot\pi - |x_i - y_i|)\}^2} \qquad (1)$$

Indeed this definition is somewhat debatable as the meaning of the distance heavily depends on dimensionality, with the largest possible distance ranging from $\sqrt{2}\pi$ to $\sqrt{6}\pi$ for the two and six dimensions (i.e., glycine vs. lysine), respectively. In practice, however, the addition of a dimension specific constant to the energy term solves this problem (see below).

Estimation of the probability density function

Let *aa* be a specific residue type, and let $\boldsymbol{X}$ denote some subset of the TAs that define the conformation of *aa* such as $\{\varphi,\psi,\chi_1\}$ or $\{\chi_1,\chi_2\}$. Let $\vec{x}$ denote a specific vector of values in the TA space of $\boldsymbol{X}$. The probability density function at $\vec{x}$ is estimated from the TA database as

$$P(\vec{x}\,|aa) = \frac{\sum_{\vec{y}} e^{\frac{-D(\vec{x},\vec{y})^2}{\sigma^2}}}{\sum_{\vec{y}}\sum_{\vec{z}} e^{\frac{-D(\vec{y},\vec{z})^2}{\sigma^2}}} \qquad (2)$$

where the summations are over all occurrences of type *aa* in the database, σ is a constant (chosen as 16° in this work), $\vec{y}$ and $\vec{z}$ are also vectors in $\boldsymbol{X}$. This formulation follows the theory of normal kernel density estimators.[40]

Basic energy terms

All the TA energy terms presented in this work (Table I) are linear combinations of basic energy terms. All the basic terms are continuous functions of up to three TAs. We refer to them as basic because they address the sets of TAs that are most strongly coupled, and thus cannot be decomposed into simpler terms. The basic terms are derived from the estimations of the probability density functions.

$$E_{aa}(\vec{x}) = -\log(P(x\,|aa)) - C_{aa} \qquad (3)$$

where C_{aa} is an empirically determined constant that depends on the number of TAs in the *aa* residue type:

Table I
A Summary of all Energy Terms Presented in this Work

Basic terms	
$Ramach2D_{aa}(\varphi, \psi)$	Energies associated with the probability of observing a given configuration of TA values in a residue of type *aa* [Eqs. (1) and (2)]. χ_i is any TA in the side-chain, apart from χ_1, that is applicable to *aa*.
$Ramach3D_{aa}(\varphi, \psi, \chi_1)$[a]	
$Chi1D_{aa}(\chi_1)$[b]	
$Chi2D_{aa}(\chi_{i-1}, \chi_i)$[b]	
Propensity terms[c]	
$EnProp2D_{aa}(\varphi, \psi)$	Energies associated with the propensity of residue type *aa* to adopt certain TA values [Eqs. (4) and (5)]
$EnProp3D_{aa}(\varphi, \psi, \chi_1)$[a]	
Composite terms[c],[d]	
$EnRESIDUE_{aa}(\varphi, \psi, \chi)$	The energy associated with the probability of observing a given configuration that encompass all the TAs that are applicable to residue type *aa* [Eq. (7)].
$EnIND_{aa}(\chi)$[b]	Same as *EnRESIDUE*, except that only the side-chain TAs are considered. This function should correlate with probabilities of backbone-independent rotamer libraries [Eq. (8)].
$EnDEP_{aa}(\varphi, \psi, \chi)$[b]	The energy associated with the probability of observing a given χ configuration in a residue of type *aa* with a certain φ and ψ backbone angles. This function should correlate with probabilities of backbone-dependent rotamer libraries (Eq. 9).

The exact computation methods are described in the text. *1D*, *2D*, and *3D* refer to the number of variables in each function.

[a] For alanine and glycine, the 3D functions are defined as equal to the 2D counterparts.

[b] These terms are defined as zero for alanine and glycine.

[c] These terms are linear combinations of basic terms.

[d] χ represents all the side-chain TAs that are applicable to type *aa*.

$C_{\text{ala,gly}} = -0.88$, $C_{\text{cys,ser,thr,val}} = -0.31$, $C_{\text{asp,phe,his,ile,leu,asn,pro,trp,tyr}} = 0$, $C_{\text{glu,met,gln}} = 0.21$, $C_{\text{lys,arg}} = 0.37$.

In low density regions, that would become high energy regions, the accuracy of the energy is sacrificed for the sake of smoothness. These regions are expected to have very little effect on the energy of the models but high energy local minima may considerably hamper simulations.

The probability density functions and the energy terms are derived by a five-step process (Fig. 1).

1. Removal of outliers [Fig. 1(A)] – The database includes rare TA combinations, whose exact contributions to the probability density functions are hard to interpret given the size of the database. Thus, they are ignored in the construction of the energy terms. An observation *(aa,x)* is ignored if fewer than 0.5% of the total observations of residue type *aa* are within a certain radius around it. This radius was chosen as 15° in the construction of 1D and 2D terms, and 25° for 3D terms. Depending on the residue type, 2.0–7.8% of the observations were excluded.
2. Grid construction [Fig. 1(A)] – The space spanned by the TAs is sampled using a grid. The denser the grid, the more accurate is the sampling, but the memory requirements also rise accordingly. In the current implementation, the $\varphi\backslash\psi$ space is uniformly sampled by an 18° grid (20 samples per TA). Side-chain TAs are sampled non-uniformly to allow wider spacing in sparsely populated areas. Each TA is sampled at eight or twelve points: at its peak density values, 18° away from them on both sides and at the midpoints between these clusters.
3. Sampling and pseudo-counts addition – First, the probability density function at the grid points is estimated [Eq. (2)] and local density maxima are identified. Then, at points with very low density (<2) we add a pseudo-count , which equals to one minus half the distance to the closest local density maximum.
4. Conversion of the density functions to energy terms [Eq. (3), Fig. 1(B,C)]. Because of the pseudo-counts no grid point has density below zero or energy above 12.92.
5. Cubic spline interpolation with periodic boundary conditions between the grid points[32] [Fig. 1(D)].

Propensity energy terms

Propensity measures the conformational preferences of a specific residue type as compared with the entire residue population.[28] It is derived from the probability density functions by:

$$\text{Propensity}_{aa}(x) = \frac{P_{aa}(x)}{P(x)} \tag{4}$$

where *P(x)* is the probability density function of the entire residue population, regardless of type. Because of the *log* function in Eq. (3), the following two- and three-dimensional propensity energy terms can be constructed from linear combinations of basic terms:

$$EnProp2D_{aa}(\varphi, \psi) = Ramach2D_{aa}(\varphi, \psi) - Ramach2D(\varphi, \psi) \tag{5}$$

$$EnProp3D_{aa}(\varphi, \psi, \chi_1) = Ramach3D_{aa}(\varphi, \psi, \chi_1) - Ramach3D(\varphi, \psi, \chi_1) \tag{6}$$

where *Ramach2D* and *Ramach3D* are type-independent functions based on the entire TA database. *EnProp3D* of alanine and glycine is set as their *EnProp2D* value.

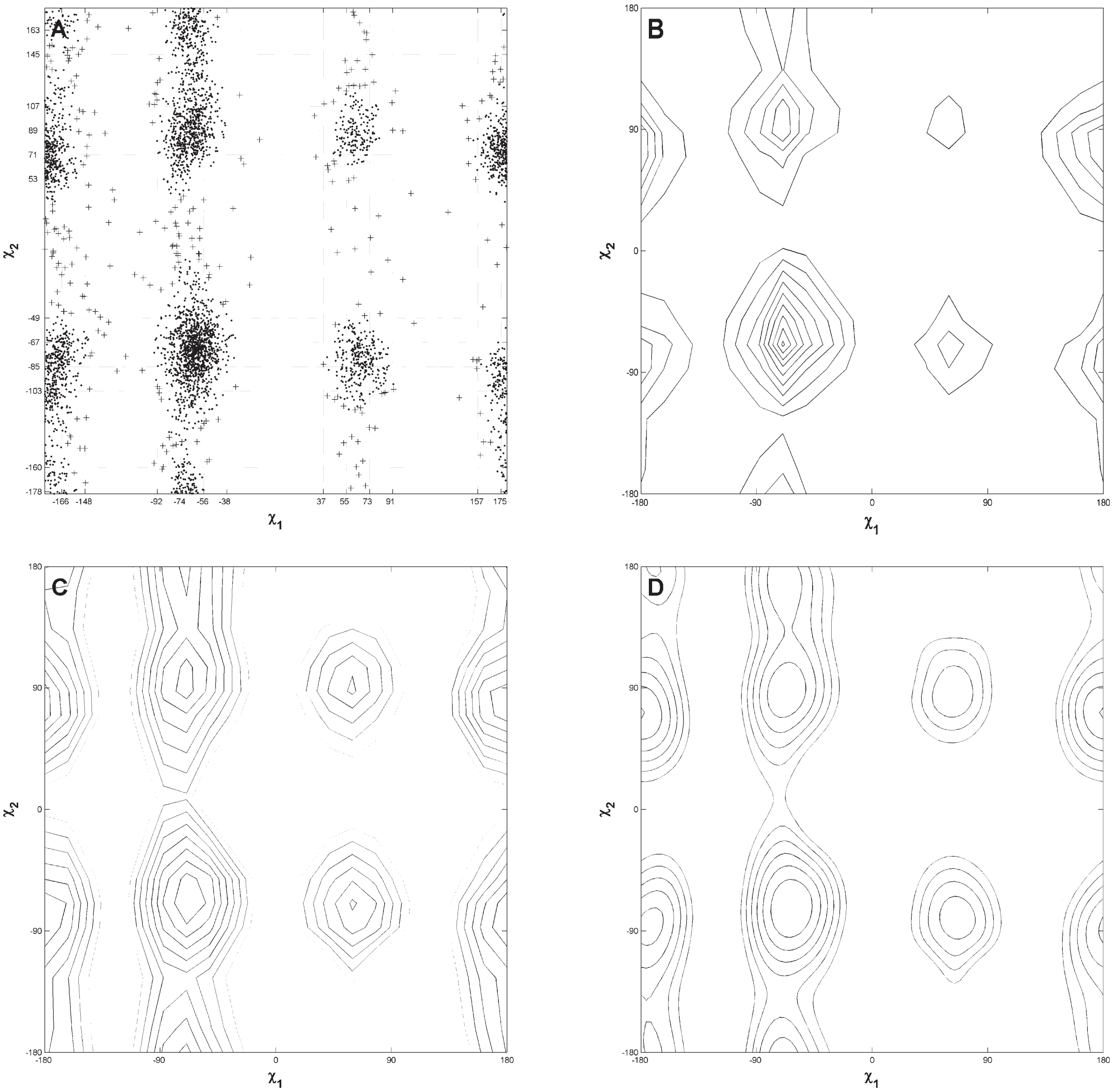

Figure 1

Construction of a basic energy term [Eqs. (2) and (3)], illustrated for the Chi2D(χ_1, χ_2) *of histidine.* ***A****. Initial TA values. Each point represents one observation of histidine χ_1 and χ_2 in the database. The crosses represent TA values of very rare conformations that are removed in later steps. The dotted lines represent the specific grid built for this amino acid type and these TAs.* ***B****. Contour plot of the probability density function at each grid point.* ***C****. Contour plot of the energy value at each grid point, after the addition of pseudo-counts.* ***D****. Contour plot of the final, two-dimensional, spline-smoothed energy term.*

Composite energy terms

This class of energy terms is associated with the probabilities of observing TA configurations of higher (up to six) dimensions. These terms are not fundamentally different from the basic ones, and could be derived directly from the database. However, because the number of coefficients that define a spline grows exponentially with its number of variables, it is impractical to use spline implementations of four variables or more with current computer hardware. Fortunately, a simplifying decomposition is possible. It relies on the observation that while χ_1 is strongly coupled to the φ and ψ TAs, the χ_2 TA is primarily coupled to χ_1, the χ_3 TA is primarily coupled to χ_2, and the χ_4 TA is primarily coupled to χ_3.[5,29] Thus the probability density function of a given TA combination may be approximated by:

$$\begin{aligned} &P(\varphi,\psi,\chi_1,\chi_2,\chi_3,\chi_4|aa) \\ &\quad \cong P(\varphi,\psi,\chi_1|aa)\cdot P(\chi_2|\chi_1)\cdot P(\chi_3|\chi_2)\cdot P(\chi_4|\chi_3) \\ &= P(\varphi,\psi,\chi_1,|aa)\cdot \frac{\Pr(\chi_2,\chi_1)}{\Pr(\chi_1)}\cdot\frac{\Pr(\chi_3,\chi_2)}{\Pr(\chi_2)}\cdot\frac{\Pr(\chi_4,\chi_3)}{\Pr(\chi_3)} \end{aligned} \qquad (7)$$

Because of the *log* function in Eq. (3), the energy associated with this probability, which we denote as *EnRESIDUE*, may be written as the following sum of basic terms:

$$EnRESIDUE_{aa}(x) = \begin{cases} Ramach2D_{aa}(\varphi,\psi) + C_{aa} & \text{if } aa \in \{A,G\} \\ Ramach3D_{aa}(\varphi,\psi,\chi_1) + C_{aa} & \text{if } aa \in \{C,S,V,T\} \\ Ramach3D_{aa}(\varphi,\psi,\chi_1) + C_{aa} & \\ \quad + \sum_{i=2}^{n}(Chi2D_{aa}(\chi_{i-1},\chi_i) - Chi1D_{aa}(\chi_i)) & \text{otherwise} \end{cases} \qquad (8)$$

where *x* is the vector (φ, ψ, $\chi_{1,..,}$ χ_n) and *n* is either 2, 3, or 4 depending on the residue type *aa*. Backbone-independent and backbone-dependent terms for side-chain TAs are likewise approximated by:

$$EnIND_{aa}(x) \cong \begin{cases} 0 & \text{if } AA \in \{A,G\} \\ Chi1D_{aa}(\chi_1) + C_{aa} & \text{if } AA \in \{C,S,V,T\} \\ Chi1D_{aa}(\chi_1) + C_{aa} \ + \sum_{i=2}^{n}(Chi2D_{aa}(\chi_{i-1},\chi_i) - Chi1D_{aa}(\chi_i)) & \text{otherwise} \end{cases} \qquad (9)$$

$$EnDEP_{aa}(x) \cong \begin{cases} 0 & \text{if } AA \in \{A,G\} \\ EnRESIDUE_{aa}(x) - Ramach2D_{aa}(\phi,\psi) & \text{otherwise} \end{cases} \qquad (10)$$

where *x* is the vector ($\chi_{1,...,}\chi_n$) for *EnIND* and (φ, ψ, $\chi_{1,..,}$ χ_n) for *EnDEP.*

Decoy sets

Four decoy sets were used in the assessment of the energy terms:

(i). The ROSETTA all-atom decoy sets (RAADS)[33] include 41 protein chains of length 35–66 residues with 1600–1900 decoys each. The native structures of 25 proteins were solved by X-ray crystallography and 16 were solved by NMR. We considered 13 of these sets as high-quality as at least 5% of their decoys were close to native (RMS $\leq$ 3.5 Å).

(ii). The LMDS decoy sets[34] include 10 protein chains of length 31–68 residues with 200–500 decoys each. The native structures of nine proteins were solved by X-ray crystallography. Only two sets are of high quality.

(iii). The CASP7 submission corpus was downloaded from the CASP7 web site, http://predictioncenter.org/casp7. Targets solved by NMR, cancelled targets and targets for which fewer than 5% of all submissions were within 3.5 Å of native were not considered. The domain parsing of the CASP7 assessors was used, resulting in 63 domains. For each domain, the submission set consists only of submissions that include all the atoms that were resolved in the native structure. The average number of valid submissions per domain is 368, and no domain has less than 250 submissions.

Native discrimination and enrichment of the decoy sets

The performance of the energy terms in native/decoys discrimination tasks was evaluated as the Z-score of the native protein energy against the distribution of decoy energies. The ability of the energy terms to assist in picking the lowest RMSD decoys was evaluated by introducing the following enrichment measure for each protein set.[33]

$$\text{Enrichment} = \frac{|\{15\%\ \text{Lowest Energy Decoys}\} \cap \{15\%\ \text{Lowest RMS Decoys}\}|}{0.15^{*}0.15^{*}|\text{Complete Decoy Set}|} \qquad (11)$$

Threading experiments

The threading experiments follow closely the protocol of Shortle.[28] A corpus of structure fragments was obtained from the 850 proteins that were used to build the TA database. All possible ungapped chain fragments with lengths of 10, 20, 30, and 40 residues were included. The number of fragments in the sets is 135,000–168,000, depending on their length. Sequences for threading are extracted from the fragment set by considering each 10th fragment, thus guaranteeing a mutual difference of at least ten residues between sequences. Each sequence is threaded through the entire fragment set of corresponding length, and the threading energy of each threading instance is evaluated by the threading potential. The energy of the sequence threaded into its native structure is then compared with the energy of all the other threading instances. If the native threading energy is within the lowest 0.1% of the incorrect threading energies we consider it a success. Threading with the *EnPROP3D* potential requires that χ_1 is defined, and this is not applicable with alanine and glycine in either the threaded sequences or the templates. In those cases the *EnPROP2D* is used instead.

Minimization experiments

The *EnRESIDUE* and *EnProp2D* energy terms were employed in a minimization experiment as a "proof of concept" of their derivability and minimization capabilities. The following energy terms were used in conjunction with them: bond length, angle, plane, out-of-plane, excluded volume and an all-atoms tether energy term. All the energy terms had equal weights. The minimization used the L-BFGS algorithm[35] and was allowed to continue until convergence.

Propensity grade

The propensity grade measures the compatibility of a protein model with the TA propensity energies observed in a large, nonredundant set of high-quality X-ray structures.

$$\mathit{PROP\_Grade} = \frac{1}{n}\sum\left(\frac{\mathit{EnPROP2D}_{aa} - \mathit{meanEnPROP2D}_{aa}}{\mathit{stdEnPROP2D}_{aa}}\right) \qquad (12)$$

Where the summation is over all the residues of the protein and $\mathit{meanEnPROP2D}_{aa}$ and $\mathit{stdEnPROP2D}_{aa}$ are the mean and standard deviations of the *EnPROP2D* energies of all the residues of type *aa* in the database, respectively. Since propensity energies are used (rather than the propensities themselves) the lower a model's propensity grade, the more closely it conforms to the database.

Implementation

Determination of the spline parameters was done in MATLAB. All other results were obtained within the framework of MESHI, a molecular modeling package written in Java.[36] Each composite or propensity term was implemented as a separate sub-package. MESHI is available for download at: http://www.cs.bgu.ac.il/~meshi. The spline parameters and the MATLAB scripts that generated them are available in MATLAB file format through the MESHI website.

RESULTS

Accuracy estimates of the new energy terms

This work provides a set of high-dimensional and differentiable TA energy terms for proteins. These terms approximate the negative log of various probability density functions (Fig. 1). Several factors may compromise the accuracy of this approximation: The relatively sparse sampling of the density function by the grid, the use of pseudo-counts, the spline interpolation and the use of linear combinations of 1D, 2D, and 3D elements to form up to six-dimensional functions. In this section we test the accuracy of three terms: *EnRESIDUE* [Eq. (8)], *EnIND* [Eq. (9)] and *EnDEP* [Eq. (10)]. These are the most error-prone terms as their derivation includes all of the approximation steps. Thus, their accuracy also implies that the basic terms are accurate. Furthermore, as the propensity terms are derived from the basic terms without any approximations [Eqs. (5) and (6)], their accuracy is also validated.

To test the accuracy of the *EnRESIDUE* term we compared it with more direct and thus more accurate (and computationally heavier) density estimation. For each residue in the database, the probability of that residue's conformation was calculated using Eq. (2). The correlation between the exponent of negative *EnRESIDUE* and the estimated probability is 0.99. This high correlation value indicates that our approximation captures much of the true distribution. A higher resolution function (using a φ/ψ grid with 9° resolution rather than 18°) achieved essentially the same results (data not shown).

The availability of rotamer libraries provides a more direct test for the side-chain terms *EnIND* and *EnDEP*. Figure 2(A) compares the probabilities of the rotamers from a backbone-independent library[30,37] to the expo-

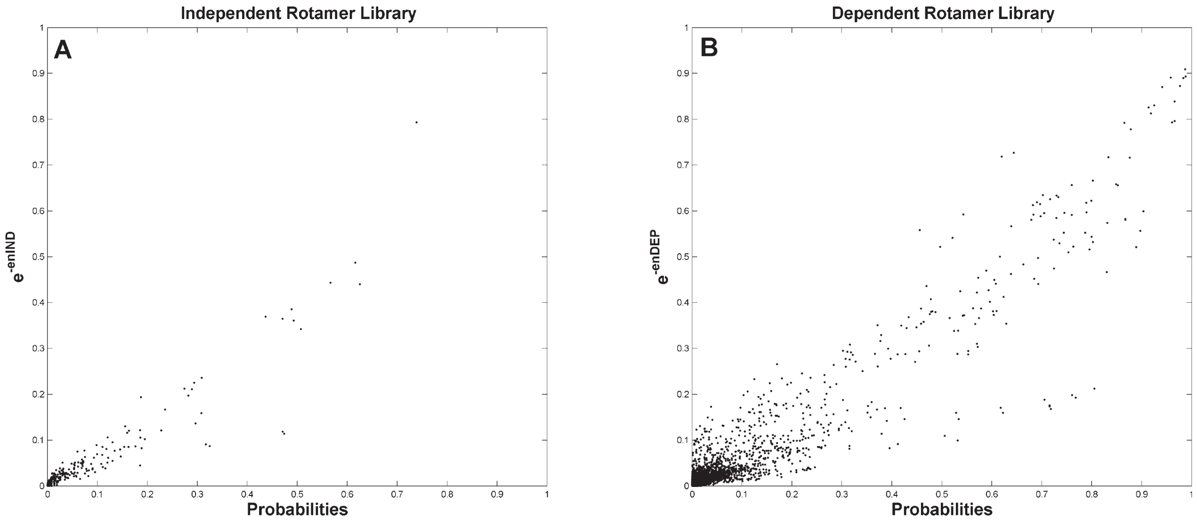

Figure 2

The correlation between the rotamer probabilities in the Dunbrack library[5,37] and our side-chain energy terms. Each point refers to one rotamer from the Dunbrack library. ***A****. The backbone independent rotamer library and the exponent of negative EnIND (correlation, R = 0.93).* ***B****. The backbone dependent rotamer library and the exponent of negative EnDEP (correlation, R = 0.93). The backbone dependent library samples the {φ,ψ} TA values with 10° bins. Only rotamers from the most occupied bins are included in the figure and correlation calculations (see text).*

nent of negative *EnIND*. The correlation between them is 0.93. Further, the more frequent rotamers often coincide with the minima of the energy term [Fig. 3(A)]. The *EnIND* term sometimes suggests continuity between rotamers [Fig. 3(B)], which is not evident from the discrete rotamer library.

The comparison between the backbone-dependent library[5] and the *EnDEP* energy term is somewhat more problematic. The backbone-dependent library estimates the rotamer probabilities for all backbone conformations including extremely rare, and perhaps impossible, ones (for example, {φ,ψ} values around {0°,0°}). Conversely,

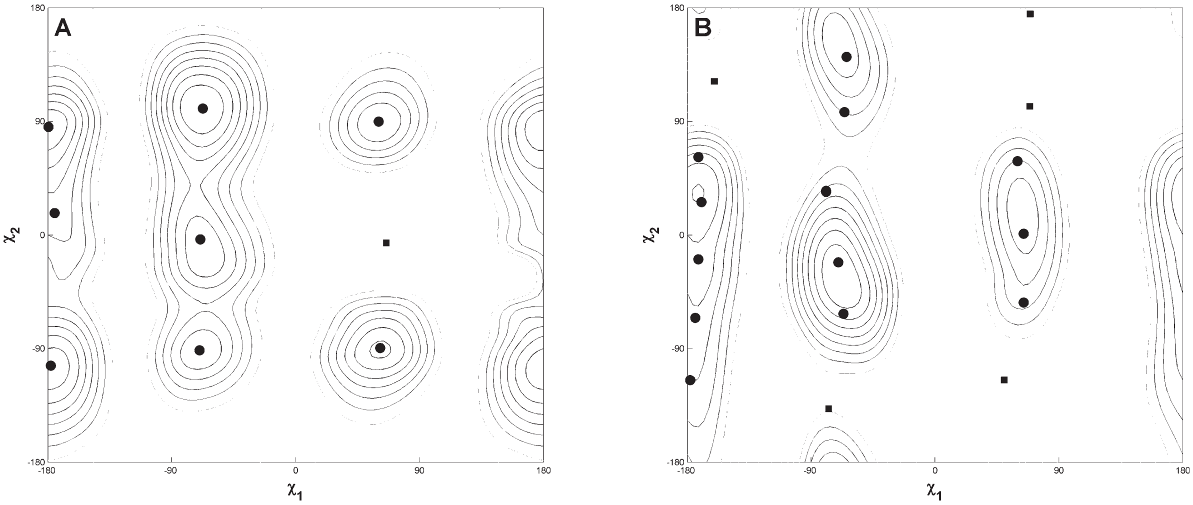

Figure 3

The contour plots of Chi2D(χ_1,χ_2) for two residue types (tryptophan (A) and asparagine (B)). Backbone-independent rotamers from the Dunbruck 2002 library are marked by circles and squares. Rotamers marked with a square appear in unfavorable regions in the plot. However, their probability according to the rotamer library is also very low (<1%). ***A****. Polynomial spline for tryptophan. Most rotamers correspond to narrow local minima.* ***B****. Polynomial spline for asparagine. The plot has several wide basins (for example, around $\chi_1 \approx -60°$, $\chi_2 \approx -75°$ to 30°). Their representation in the spline is continuous, but a discrete rotamer library is forced to divide them to several different rotamers.*

our pseudo count scheme was mainly designed to result in a high and smooth energy landscape in these conformations. There are no high energy local minima that may be attributed to the rotamers. Thus, when all the rotamers are considered, there is no correlation between the rotamer probabilities and the exponent of negative *EnDEP* (data not shown). On the other hand, when only the most occupied regions of the Ramachandran plot are considered (10° × 10° bins that include at least 5% of the database observations), the correlation is 0.93 [Fig. 2(B)].

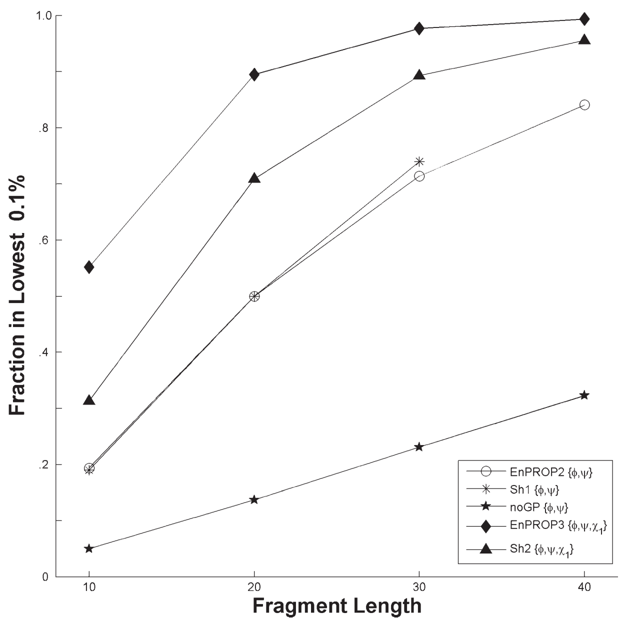

Figure 4

The fraction of 16,000 sequences for which the native conformation ranked in the lowest 0.1% of conformations (170 per 170,000) by propensity-based threading potentials. Each line is the result of a different potential. Sh1 data are taken from Fang and Shortle,[38] Sh2 data are taken from Shortle[28] and "noGP" data are calculated by EnPROP2D after the exclusion of glycine and proline contributions.

Threading experiments

Another indirect evidence for the accuracy of the energy terms is their ability to reproduce and improve on previously reported results. Kocher *et al.*[27] and later Fang and Shortle[28,38] showed that backbone conformational propensities are very effective in ungapped threading of protein fragments. We repeated and expanded these experiments using the *Ramach2D*, *EnPROP2D* and *EnPROP3D* terms. More than 16,000 sequence fragments of lengths 10, 20, 30, and 40 amino acids were threaded through about 170,000 structure fragments of corresponding lengths. The fraction of sequences with energies in the lowest 0.1% range (success rate) is shown in Figure 4, as a function of the fragment length. The results are consistent with the studies of Fang and Shortle, and show that the propensity term very accurately identifies the native threading when the sequences are long. The success rates of *EnPROP2D* are 19.4%, 50%, 71.4%, and 84.1% for fragments of lengths 10, 20, 30, and 40, respectively. These numbers are nearly identical to the results obtained by Fang and Shortle[38] that used a discrete {φ,ψ} propensity potential that was based on 10° × 10° binning of the Ramachandran plot. The success rates of *EnPROP3* are 55.2%, 89.5%, 97.7%, and 99.4% for fragments of lengths 10, 20, 30, and 40, respectively. These rates are higher than those reported by Shortle,[28] probably because of the finer grid sampling that was used to construct *EnPROP3D*. In accord with previous studies, we find that the conformational probabilities, represented by our *Ramach2D* term, are much less suitable than propensities as threading potentials. The use of *Ramach2D* as the threading potential resulted in low success rates of 0.15%, 0.1%, 0.1%, and 0.15% for fragments of lengths 10, 20, 30, and 40, respectively.

We further studied the specific role of the two exceptional residue types, proline and glycine, in the success of the propensity terms at picking native structure needles from a huge haystack of non-native fragments. Because of their special structures, propensities of glycines and prolines have extreme values that are unmatched in the other eighteen types. For example, threading of proline into a structure that is forbidden by the proline stereochemistry will result in very high propensity values. Likewise, threading of any residue into a structure that originated from a glycine-specific conformation will lead to similar high propensity values. Thus, one may suggest that the propensity terms merely identify correct registration of glycines and prolines between sequence and structure. We checked this claim with another threading experiment, termed "noGP," in which the contributions of prolines and glycines were fully excluded. To this aim, the propensities of prolines and glycines in the sequence were not added to the total propensity energy. We also discarded propensities of residues that were threaded into structures that originated from prolines or glycines. The results of this experiment (Fig. 4) show a marked decrease in the success rates to 5.0%, 13.7%, 23.1%, and 32.3% for fragments of lengths 10, 20, 30, and 40, respectively. These results indicate that glycines and prolines are indeed responsible for the large part of the success rates of the original experiments. Still, the success rates of the "noGP" experiment are considerable even at short fragment lengths, suggesting that the propensity provides a genuine signal for the fit between sequence and structure.

Identification of native and native-like structures within decoy sets

Native energy Z-score and enrichment are two common benchmarks in assessing the performance of energy

Table II
Mean Z-scores of the Native Structure Energies

Decoy set	*EnRESIDUE*	*EnPROP2D*
LMDS (10 domains)	−1.33	−4.24
LMDS, w/o NMR (9 domains)	−2.31	−4.55
RAADS (41 domains)	2.8	−1.12
RAADS, w/o NMR (25 domains)	−0.06	−2.37
CASP7 Corpus (63 domains)	−0.5	−1.89

The lower the Z-score, the better the energy discrimination between the native structures and the decoys. NMR models yielded extremely high *EnRESIDUE* and *EnPROP2D* energy values, and the scores improve when they are removed from the complete decoy sets.

terms.[33] The native energy Z-score quantifies the confidence level in energy-based identification of native structure within a large decoy set. Enrichment quantifies the ability to assess model quality. High enrichment values indicate that the energy term is able to identify the best models within the decoy set.

Here we tested two of the energy terms, *EnPROP2D* and *EnRESIDUE*, using three decoy sets. In the Z-score test (Table II), the *EnPROP2D* term is clearly superior to the *EnRESIDUE* term on all sets, in agreement with the threading experiments. Both terms, however, perform worse when the native structure is solved by NMR, and the Z-scores improve considerably when the NMR-containing sets are removed from the test. This is not surprising, as deviations of TA statistics in NMR structures from those observed in X-ray structures were already reported.[13–15] Yet, the ability of our energy terms to detect these deviations is encouraging.

Of special interest are the Z-score results obtained by the *EnPROP2D* on the CASP7 submission corpus. A-priori, it was expected to be the most difficult decoy set as we considered only targets for which good templates were available during the prediction event and the models were created by the leading prediction groups. Yet, the average Z-score of *EnPROP2D* on this set is remarkably low. Moreover, targets in which the native structures had considerably lower propensity than all the submissions were quite common. For example, target T0328 [Fig. 5(A)] was designated by the CASP7 assessors as high-accuracy target, and indeed most of its submissions have a 1.5–3 Å RMSD from native structure. Despite that, its *EnPROP2D* energy is lower than that of any submission. In all, the *EnPROP2D* energy of the native structure was ranked as lowest in 37% of the domains and was ranked among the lowest five submissions in 57% of the domains [Fig. 5(B)].

Analysis of the submission propensities of specific groups in the CASP7 decoy set show that group 020 (Baker) was unique in two ways. First, it submitted 100 models that had lower propensity energies (and thus higher propensities) than the corresponding native structure. This is in great contrast to the next group by this criterion (group 556 with 32 submissions) or the baseline (about 20 submissions per group). Second, in many cases it was the *only* group to submit models with lower *EnPROP2D* energy than native [Fig 5(B)]. The ROSETTA package that was developed by the Baker group uses both TA propensity and probability terms in its forcefield.[22] The above results indicate that these terms fulfilled their purpose. The effect of these terms is also

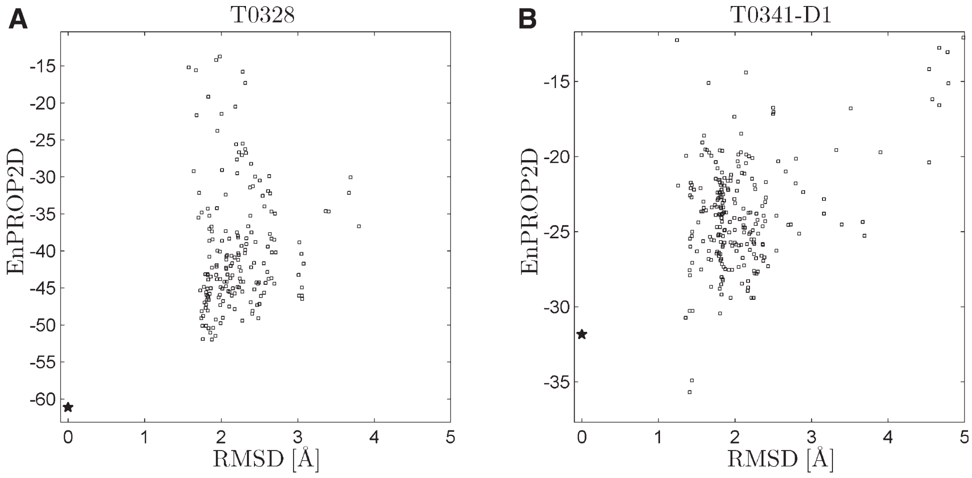

Figure 5

*The values of EnPROP2D as a function of RMSD from native for the submissions of two CASP7 targets: (**A**) T0328 (306 residues) and (**B**) T0341 domain 1 (148 residues). Submissions are marked with squares and natives are marked with stars. The three lowest propensity models in the T0341 submissions were from group 020 (Baker).*

Table III
Mean Enrichment of the Decoy Set in Close-to-Native-Structures [Eq. (10)]

Decoy set	*EnRESIDUE*	*EnPROP2D*
LMDS (10 domains)	1.25	1.02
RAADS (41 domains)	1.12	1.28
RAADS, High Quality (13 domains)	1.59	1.53
CASP7 Corpus (63 domains)	1.62	1.54

The high quality RAADS decoy set and the CASP7 corpus include relatively many decoys similar to the native. Their high scores indicate that the terms can choose the more native like structures from a set of good structures and are appropriate for structure refinement.

noticed in the ROSETTA generated RAADS set, which proves the most difficult set in the Z-score test for both *EnPROP2D* and *EnRESIDUE*. This difficulty is in great contrast to the LMDS set that was derived with a very naïve treatment of TA.

Enrichment experiments, which test the ability of energy terms to pick the best models out of the decoy sets, show a somewhat different picture (Table III). In this test, the probability and the propensity energy terms show similar performance. Both perform well on the high-quality decoy sets that include a large fraction of native-like decoys. The enrichment of the other sets degrades as the sets become less native-like. Still, enrichment values above random selection exist also in the LMDS set. The study that introduced the RAADS set[33] also tested the enrichment of more than 20 energy terms on the 41 sets (including solvation terms, hydrogen-bonding terms and so on). The highest enrichment value that was observed for a single term was 1.61, and only seven of them had enrichment above 1.3. On that scale, the enrichment values of both terms on the RAADS set are fairly good.

Protein structure minimization

The motivation behind the protein structure minimization experiment is twofold. The first goal is technical, to verify that minimization of a model under a forcefield that includes the TA energy terms can indeed lead to improvement in the TA criteria (i.e., the energy terms do what they were designed to do). A second, more general goal is to improve the quality of NMR structures. NMR structures often perform poorly according to standard TA criteria,[13–15] as was also shown by the decoy set tests in this work. Thus, improving the local quality of NMR-derived structures without considerable global changes may serve as an interesting test to the usefulness of our terms.

The thirteen NMR structures of the RAADS set perform poorly in the TA test (CHICHK) of the WHATIF package.[6] Their average Z-scores for $\{\varphi,\psi\}$ and $\{\chi_1,\chi_2\}$ distributions are -4.2 and -3.9, respectively, while Z-scores below -3 are considered "worrying." Only three of these structures have $\{\varphi,\psi\}$ Z-scores above -3 and all of them are in the "poor" range between -2 and -3. Their mean propensity grade is 0.15.

The native structures were energy minimized under an energy function that included the *EnRESIDUE* and *EnPROP2D* terms (see methods). The average RMSD of the minimized models from the original structures is only 0.5 Å. This slight global shift is accompanied by a considerable improvement in the results of local quality tests. The average $\{\varphi,\psi\}$ and $\{\chi_1,\chi_2\}$ Z-scores of the minimized models are 2.477 and 1.682, respectively, and none of them were "poor" according to the CHICHK definitions. Their mean propensity grade has improved to -0.07. In the control experiment, the minimization was performed without the *EnRESIDUE* and *EnPROP2D* terms. The average RMSD of the minimized models from the original structures was 0.24, lower than the experiment group, and yet, the average $\{\varphi,\psi\}$ and $\{\chi_1,\chi_2\}$ Z-scores decreased to -4.43 and -4.3, respectively, and the mean propensity grade was 0.14.

Testing the compatibility of the minimized structures with the NMR constraints is beyond the scope of the current work. Thus, we do not claim that these structures are better than the initial structures in some biologically relevant way. Still, we believe that these results suggest that the new terms may be useful in this context.

DISCUSSION

Continuous first derivatives are a desired and often required property of energy terms in force-driven algorithms. Two approaches have been suggested so far for the development of knowledge based TA energy terms with this property. The first approach is to represent the TA energy landscape by a sum of Gaussian functions[17,19,21] or Fourier terms.[39] Although any functional form may be approximated by such a sum, high accuracy representations of nonregular landscapes require the summation of a large number of terms. Thus, a trade-off exists between accuracy and evaluation speed for this approach. We chose an alternative, spline-based, approach, which is similar to the one used by Fujitsuka *et al.*[26] With this approach the CPU requirements per TA are constant regardless of the accuracy. However, the number of spline coefficients grows alarmingly fast with the number of dimensions in the function, and splines with six TA variables (the number of coupled TAs in lysines or arginines) are currently beyond the memory capacities of most computers (several Gigabytes). We circumvent this problem by using low-density and nonuniform sampling and also by approximating high-dimensional functions by linear combinations of low-dimensional ones. A major result of the current study is that these approximations do not considerably compromise accuracy. In fact we show that the resulting energy terms are compatible with the criteria set by widely used structure assessment tools such as PROCHECK or WHATIF. Thus, the current work presents, for the first time, an accurate, computationally efficient and fully differentiable treatment of the full TA space of any residue. We utilize this approach to derive a series of terms that describe the probability density functions of backbone conformations, side-chain conformations and the conformations of entire residues. We also derive terms that describe the propensity of residue types to adopt certain backbone conformations (Table I).

Our emphasis on accuracy is relaxed in the case of underpopulated conformations. They are very sparsely sampled and further, their observed densities were modified by the pseudo counts. Indeed, our major concern regarding these conformations was to avoid high energy local minima. We believe that the resulted inaccuracy is negligible as these conformations are hardly or not at all represented in native structures. High energy local minima, on the other hand, may slow the convergence of optimization experiments. It should be noted that this decision is not inherent to the approach presented here. A different pseudo-count scheme would result in higher accuracy in the rare conformations.

From the native discrimination test that we performed on the CASP7 submission corpus (Table II and Fig. 5), it is apparent that the class of terms described in this study is not yet used to its full strength. Some of the most advanced forcefields available contributed to this decoy set, and still the majority of them failed to reproduce the native propensity energy (*EnProp2D*). The situation is better with regard to the probability energy (*EnRESIDUE*), but even here the average Z-score is negative, which means that many prediction methods reached too high probability energy. Similarly, TA propensity terms are hardly ever used in the energy annealing stage of NMR structure determination, resulting in structures with higher average propensity energy than X-ray structures. Whether the reasons for the TA terms being underused are concerns about accuracy, speed or differentiability, we hope that this work will facilitate their use.

Although we believe that energy terms like the ones presented here are important ingredients in protein structure prediction schemes, they clearly do not encapsulate all the complexity of protein TAs. Probability-based energies (e.g. *EnRESIDUE*) of many decoys are lower than those of the native structures. Furthermore, models that are subjected to minimization runs with a simplistic energy function that only include bonded and TA terms almost always reach a final state with probability and propensity energies that are much lower than native (data not shown). Clearly, other physical forces such as solvation and hydrogen bonding strongly influence the TAs. The optimal weighting of different energy terms is an application-specific problem. It is not obvious that the same weighting is optimal for structure optimization and model selection. Thus, we did not address weighting in this general work. It should be noted, though, that in the minimization experiment we received reasonable results with naïve equal weighting of all energy terms.

ACKNOWLEDGMENTS

N.K. was supported by the Kreitman Foundation Fellowship. C.K. is a Ralph Selig Career Development Chair in Information Theory.

REFERENCES

1. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–99.
2. Janin J, Wodak S, Levitt M, Maigret B. Conformation of amino acid side-chains in proteins. J Mol Biol 1978;125:358–386.
3. Bhat TN, Sasisekheren V, Vijayan M. An analysis of side chain conformations in proteins. Int J Pept Protein Res 1979;13:170–184.
4. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
5. Dunbrack RL, Jr, Karplus M. Backbone dependent rotamer library for protein application to side-chain prediction. J Mol Biol 1993; 230:543–574.
6. Kleywegt GJ, Jones TA. Phi/Psi-chology: Ramachandran revisited. Structure 1996;4:1395–1400.
7. Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 1996;5:1067–1080.
8. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
9. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992;12: 345–364.
10. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
11. Lovell SC, Davis IW, Arendall WB, III, Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins 2003;50: 437–450.
12. Sims GE, Kim SH. A method for evaluating the structural quality of protein models by using higher-order φ-ψ pairs scoring. Proc Natl Acad Sci USA 2006;103:4428–4432.
13. Heringa J, Argos P. Strain in protein structures as viewed through nonro-tameric side chains. II. Effects upon ligand binding. Proteins 1999;37:44–55.
14. Clore GM, Gronenborn AM. Comparison of the solution nuclear magnetic resonance and X-ray crystal structures of human recombinant interleukin-1β. J Mol Biol 1991;221:47–53.
15. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996; 8:477–486.
16. Spronk CA, Linge JP, Hilbers CW, Vuister GW. Improving the quality of protein structures derived by NMR spectroscopy. J Biomol NMR 2002;22:281–289.
17. Kuszewski J, Clore GM. Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. J Magn Reson 2000;146:249–254.
18. Bertini I, Cavallaro G, Luchinat C, Poli I. A use of Ramachandran potentials in protein solution structure determinations. J Biomol NMR 2003;26:355–366.

19. Levitt M. Molecular dynamics of native protein. I. Computer simulation of trajectories. J Mol Biol 1983;168:595–620.
20. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
21. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
22. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
23. MacKerell AD, Jr. Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 2004;25:1584–1604.
24. MacKerell AD, Jr, Feig M, Brooks CL, III. Improved treatment of the protein backbone in empirical force fields. J Am Chem Soc 2004;126:698–699.
25. Pohl FM. Empirical protein energy maps. Nat New Biol 1971;234: 277–279.
26. Fujitsuka Y, Chikenji G, Takada S. SimFold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins 2006;62:381–398.
27. Kocher JP, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
28. Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11:18–26.
29. Lovell SC, Michael Word J, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
30. Dunbrack RL, Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440.
31. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
32. de Boor C. A practical guide to splines. Applied Mathematical Sciences, New York: Springer; 1978.
33. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76–87.
34. Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003;329:159–174.
35. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45:503–528.
36. Kalisman N, Levi A, Maximova T, Reshef D, Zafriri-Lynn S, Gleyzer Y, Keasar C. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 2005;21:3931–3932.
37. Dunbrack RL, Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
38. Fang Q, Shortle D. Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments. Proteins 2003;53 (Suppl 6):486–490.
39. Pertsemlidis A, Zelinka J, Fondon JW, III, Henderson RK, Otwinowski Z. Bayesian statistical studies of the Ramachandran distribution. Stat Appl Genet Mol Biol 2005;4:Article35.
40. Silverman BW. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, London: Chapman & Hall/CRC; 1986.