Improvement of Statistical Potentials and Threading Score Functions Using Information Maximization

Armando D. Solis and S. Rackovsky*
Department of Pharmacology and Biological Chemistry, Mount Sinai School of Medicine, Box 1215, One Gustave L. Levy Place, New York, New York 10029

ABSTRACT **We show that statistical potentials and threading score functions, derived from finite data sets, are informatic functions, and that their performance depends on the manner in which data are classified and compressed. The choice of sequence and structural parameters affects estimates of the conditional probabilities $P(C|S)$, the quantification of the effect of sequence S on conformation C, and determines the amount of information extracted from the data set, as measured by information gain. The mathematical link between information gain and mean conformational energy, established in this work using the local backbone potential as model, demonstrates that manipulation of descriptive parameters also alters the "energy" values assigned to native conformation and to decoy structures in the test pool, and consequently, the performance of such statistical potential functions in fold recognition exercises. We show that sequence and structural partitions that maximize information gain also minimize the mean energy of the ensemble of native conformations. Moreover, we establish an informatic basis for the placement of the native score within an energy spectrum given by the decoy pool in a threading exercise. We discover that, among all informatic quantities, information gain is the best predictor of threading success, even better than the standard Z-score. Consequently, the choices of sequence and structural descriptors, extent of compression, and levels of discretization that maximize information gain must also produce the best potential functions. Strategies to optimize these parameters with respect to information extraction are therefore relevant to building better statistical potentials. Last, we demonstrate that the backbone torsion potential, defined by the trimer sequence, can be an effective tool in greatly reducing the set of possible conformations from a vast decoy pool. Proteins 2006;62:892–908.** **© 2006 Wiley-Liss, Inc.**

Key words: statistical potentials; local potential; threading; information theory; information gain; protein structure

INTRODUCTION

Computational protein structure analysis and prediction rely heavily on the accurate estimation of sequence-dependent structural probability distributions from values of a number of specific properties, including amino acid sequence and protein environment. These probabilities are used to detect significant sequence-structure patterns,[1–3] to guide computational searches for native structure from sequence information,[4–7] as components of empirical force fields or potentials of mean force,[8–12] and as scoring functions for threading and structure recognition methods.[13–17] Although these applications rely on estimating probability distributions from experimental data, only a few studies have been directed specifically to optimizing their performance.[18–24]

Statistical potentials, which are based on these probability distributions, attempt to describe forces important in stabilizing the native structure of protein chains. In this work, we demonstrate a fundamental mathematical relationship between these potentials and various informatic quantities. This direct correspondence allows us to treat "energies" derived from database statistics as purely informatic functions, without invoking any biophysical context, and eliminates the need to reconcile them with real force fields. It also makes them directly amenable to various methodologies already available to optimize information gain.

We have shown previously that the way sequence and structure data are represented and compressed affects our ability to effectively estimate conditional probabilities, and therefore the amount of useful information available, from a database of finite size.[2,3] In the present work, we demonstrate that statistical potentials derived from probability functions with maximal information gain significantly outperform other potentials in the task of identifying native conformations of protein chains. Through such quantitative connections, algorithms designed to optimize the performance of conditional probability distributions can become a basis for strategies to build better empirical potentials.

We use the effect of trimer sequence on backbone dihedral angle distributions as a concrete model of the relationship between informatics and energetic score functions. This interaction is frequently included in generalized energy functions as a standard local interaction

*Correspondence to: S. Rackovsky, Department of Biomathematical Sciences, Mount Sinai School of Medicine, Box 1023, One Gustave L. Levy Place, New York, NY 10029. E-mail: shelly@camelot.mssm.edu

Received 13 November 2004; Accepted 12 January 2005

Published online 4 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20501

component, usually termed the short-range or backbone dihedral angle potential (e.g., refs. 9 and 25). We formulate an optimized backbone torsional potential, utilizing local trimer sequence, which we show to significantly outperform single-residue potentials in threading exercises.

THEORY

Boltzmann Formalism

We begin our discussion of the informatic nature of empirical potentials with an examination of their computational foundations. We first summarize the major principles involved in building such potentials from experimental data (see ref. 14 for further detail). The Boltzmann principle, the basis of empirically derived energy functions, establishes a connection between the energies of conformational states and their corresponding probabilities of occurrence (at equilibrium):

$$p(c_i) = \exp(-E(c_i)/kT) / \textstyle\sum_j \exp(-E(c_j)/kT) \quad (1a)$$

where k is Boltzmann's constant, T is the absolute temperature, i refers to the conformational state of interest, and the summation j runs over all allowed states of the system. The denominator, $Z(C) = \sum_j \exp(-E(c_j)/kT)$, is the partition function or Boltzmann sum. The inverse Boltzmann law facilitates the assignment of an energy value to any state, given a probability density or distribution function $p(c)$,

$$E(c_i) = -kT \ln p(c_i) + kT \ln Z(C) \quad (1b)$$

(We follow the convention of using upper case letters to signify variables, and lower case letters to represent particular instances of those variable.) Such potentials of mean force are useful in determining the energy difference brought about by a specific interaction. Indicating the probability of a conformation state c_i in the presence of a specific interaction s_k as $p(c_i|s_k)$, the energy difference or net energy due to s_k is

$$\Delta E(c_i|s_k) = E(c_i|s_k) - E(c_i) = -kT \ln [p(c_i|s_k)/p(c_i)] + kT \ln [Z(c)/Z(c|s)] \quad (2a)$$

In the case of a nonlocal side-chain pair interaction, for example, c_i may refer to the distance between side chains and s_k to the amino acid identities of the side chains involved. When considering local-sequence effects on backbone torsions, c_i becomes the (ϕ,ψ) dihedral angle pair and s_k is the local amino acid sequence surrounding the backbone site of interest. The usual assumption[8,14] is to set $Z(C) = Z(C|S)$, simplifying the potential as

$$\Delta E(c_i|s_k) = E(c_i|s_k) - E(c_i) = -kT \ln [p(c_i|s_k)/p(c_i)] \quad (2b)$$

Database statistics are utilized to estimate the probabilities involved in the calculation of these potentials, by subdividing the conformational space and constructing a histogram of frequencies from a nonredundant data set. The success of this simple formalism in many applications belies its weak biophysical underpinnings.[26]

Information Gain and Average Net Potentials

The connection between informatic quantities and statistical potentials is reflected in their obvious mathematical similarity. The net energy difference averaged over all c_i and s_k is

$$\langle \Delta E(C|S) \rangle = \textstyle\sum_k \sum_i \{-kT \ln [p(c_i|s_k)/p(c_i)]\} p(c_i,s_k) \quad (3a)$$

which, upon expansion, becomes

$$\langle \Delta E(C|S) \rangle = -kT \textstyle\sum_k p(s_k) \sum_i p(c_i|s_k) \ln p(c_i|s_k) + kT \sum_i \ln p(c_i) \sum_k p(c_i,s_k) = kT \sum_k p(s_k) H(C|s_k) + kT \sum_i p(c_i) \ln p(c_i) = kT[H(C|S) - H(C)] \quad (3b)$$

where the entropies have their usual definitions[3]

$$H(C) = -\textstyle\sum_i p(c_i) \ln p(c_i)$$

and

$$H(C|s_k) = -\textstyle\sum_i p(c_i|s_k) \ln p(c_i|s_k)$$

Because

$$I_g(S,C) = H(C) - H(C|S) \quad (3c)$$

(the information gain between the two variables C and S), the average net energy is

$$\langle \Delta E(C|S) \rangle = -kT\, I_g(S,C) \quad (3d)$$

In previous work, we examined ways to optimize the impact of local sequence knowledge (S) on knowledge of backbone conformation (C) by using $I_g(S,C)$ as an objective function.[2,3] The relation above makes it clear that the average net potential is also dependent on the way the variables S and C are defined and discretized, and the way the probabilities are derived from database statistics. Both $I_g(S,C)$ and $\Delta E(c_i|s_k)$ are dependent on estimates of $p(c_i|s_k)$, and therefore all analyses pertaining to the maximization of $I_g(S,C)$ are relevant to optimization of $\langle \Delta E(C|S) \rangle$.

Detailed database-derived probabilities $p(c_i|s_k)$ that maximize $I_g(S,C)$ also minimize $\langle \Delta E(C|S) \rangle$. In this work, we estimate $p(c_i|s_k)$ from database statistics using an optimized combination of sequence-specific and background distributions, a procedure we formulated in previous work.[3] The $p(c_i|s_k)$ which maximize $I_g(S,C)$ are then used to minimize the average net energy contribution of a specific interaction.

The equivalence expressed in Eq. (3d) displays the arbitrary nature of statistical potentials. To correctly predict native structure, empirical potentials ideally should resemble actual interaction potentials describing the energy landscape of the protein chain. But the physicality of statistical potentials is difficult to prove. Reasons for this difficulty include the fact that there are many possible levels of parameterization of both conformational and sequence states that can be used to construct these potentials, and second, the statistical pressure exerted by limited data sets impairs accurate estimation of $p(c_i|s_k)$. Nonetheless, our analysis here suggests that empirical potentials are still effective in protein structure prediction because they are true informatic functions, even though their relationship to actual energies may be tenuous.

Sequence-Structure Alignment and Threading Score

Threading or folding recognition involves the proper matching of a query one-dimensional sequence to its native three-dimensional structure, using a procedure designed to discriminate true conformations from incorrect folds.[13] Threading requires two basic components: a scoring function to evaluate the fitness of any given conformation, and a large pool of decoy conformations which, one hopes, includes the native conformation.[27] A good scoring potential should consistently assign the best score to the native conformation, rather than one of the incorrect folds in the pool. Many scoring potentials are empirical energy functions, for instance, database-derived statistical potentials of the type in Eq. (2), and therefore the best score means the lowest "energy." The pool of conformations may be a set of experimentally determined protein structures or a set of artificially generated structures with native-like characteristics.

To demonstrate the informatic basis of the threading procedure, we derive a mathematical relationship between threading potential functions and an informatic quantity called divergence. We study local sequence-dependent backbone torsional potentials, and demonstrate that probability distributions giving higher information gain are likely to perform better in threading exercises.

We test the threading performance of backbone potentials on relatively short segments of amino acid sequences and their observed backbone conformations. From a data set of experimental structures, one can organize a set of short segments much larger than the set of complete protein chains, and provide a more diverse pool of decoy conformations.

The match between a sequence $s_j = s_1 s_2 \ldots$ and a putative conformation $c_k = c_1 c_2 \ldots$ may be scored, per interaction site (s', c_i), via the function

$$\Delta E(c_i|s') = -\ln\,[p(c_i|s')/p(c_i)] \tag{4}$$

where s' is a subset of s_j. For instance, if we consider the effect of local trimer sequence on backbone conformation, $s' = s_{j-1}, s_j, s_{j+1}$ and $c_j = (\phi_j, \psi_j)$. (This function is typically multiplied by kT to bring it to scale with other energies. However, because this factor does not affect the relative energy values, we drop it for simplicity.) These individual scoring terms are summed to arrive at a total energy for threading s_i on c_k

$$\Delta E_m(c_k|s_i) = (1/m) \sum\nolimits_k^m \Delta E(c_k|s_i) = -(1/m) \sum\nolimits_k^m \ln\,[p(c_k|s')/p(c_k)] \tag{5}$$

where m is the number of points of interaction in the chain. (Normalizing energy by m conveniently retains informatic relationships without altering relative rankings.)

Each c_k in the test pool of conformations gives a characteristic $\Delta E_m(c_k|s_i)$. The strength of the $\{s_i, c_k\}$ match is evaluated by comparing $\Delta E_m(c_k|s_i)$ to $\langle \Delta E_m(C|s_i) \rangle$, the average energy, per interaction site, of sequence s_i threaded onto all the conformations in the test pool. This is

$$\langle \Delta E_m(C|s_i) \rangle = -\sum\nolimits_k^r p(c_k) \Delta E_m(c_k|s_i). \tag{6a}$$

where $p(c_k)$ is the probability of occurrence of conformation c_k, and r is the number of unique conformations in the pool. If the conformation set comes from a nonredundant set of proteins, a simple average is a sufficient estimate:

$$\langle \Delta E_m(C|s_i) \rangle = -(1/n) \sum\nolimits_k^n \Delta E_m(c_k|s_i) \tag{6b}$$

where n is the number of structures in the nonredundant pool. A Z-score is computed to quantify the magnitude of the comparison

$$Z(s_i, c_k) = [\Delta E_m(c_k|s_i) - \langle \Delta E_m(C|s_i) \rangle] \,/\, \sigma(s_i, C) \tag{7}$$

where $\sigma(s_i, C)$ is the standard deviation of the spectrum of energies given by the test pool. The success of any probability distribution $p(c|s)$ in identifying the true conformation c^T for a given sequence s_i can be measured by the distance between the energy $\Delta E_m(c^T|s_i)$ of the native or correct structure c^T given by $p(c|s)$ and its associated $\langle \Delta E_m(C|s_i) \rangle$, the spectral mean. Because lower (negative) energies are favorable, a *negative* Z-score is desirable when measuring the energy of c^T against the pool.

Divergence, Statistical Potentials, and Threading Score

The relationship between threading functions and fundamental informatic quantities occurs through an informatic quantity called directed divergence (also referred to as the Kullback-Leibler distance or relative entropy), a "distance" between two distributions,

$$D(Q\|R) = \sum\nolimits_i q_i \ln\,[q_i/r_i] \tag{8}$$

where Q is the base distribution, R is the test distribution, and q and r are their specific instances.[28] (There are many interpretations of the nature of the base and test distributions, depending on the application. Our use of these terms will become clear as we proceed.) Note that $D(Q\|R) \neq D(R\|Q)$, which disqualifies D from being a metric. The term "directed" is therefore used to emphasize its directionality. A useful property of divergence is that $D(Q\|R) \geq 0$, with equality only when $Q = R$. In essence, the directed divergence measures the "error," or the increase in entropy, caused by using the "incorrect" distribution R instead of the "correct" distibution Q. The total divergence

$$J(Q\|R) = D(Q\|R) + D(R\|Q) \tag{9}$$

removes the distinction between Q and R because $J(Q\|R) = J(R\|Q)$. [Sometimes the sum $D(Q\|R) + \mathrm{D}(R\|Q)$ is normalized by 2, transforming the quantity into an average divergence.) Because $D(Q\|R) \geq 0$, $J(Q\|R) \geq 0$, with the same condition for equality.

We first demonstrate a relationship between directed divergence and information gain. The divergence between the joint distribution of C (conformation) and S (sequence) and the product of their individual distributions, $D[(C,S)\|(C)(S)]$, quantifies the informatic "distance" generated by considering the effect of sequence on conformations:

$$\begin{aligned} D[(C,S)\|(C)(S)] &= \textstyle\sum_{(c,s)} p(c,s) \ln [p(c,s)/p(c)p(s)] \\ &= \textstyle\sum_{(c,s)} p(c,s) \ln [p(s)p(c|s)/p(c)p(s)] = \sum_{(c,s)} p(c,s) \ln p(c|s) \\ &\quad - \textstyle\sum_{(c,s)} p(c,s) \ln p(c) = \sum_{s} p(s) \sum_{c} p(c|s) \ln p(c|s) \\ &\quad - \textstyle\sum_{c} p(c) \ln p(c) \sum_{s} p(s|c) = -\sum_{s} p(s) H(C|S = s) \\ &\quad - \textstyle\sum_{c} p(c) \ln p(c) = -H(C|S) + H(C) \end{aligned} \quad (10a)$$

from which follows

$$D[(C,S)\|(C)(S)] = I_g(C,S) = -\langle \Delta E(C|S) \rangle \quad (10b)$$

via Eq. (3d) (ignoring kT). Similarly:

$$\begin{aligned} D[(C)(S)\|(C,S)] &= \textstyle\sum_{(c,s)} p(c)p(s) \ln [p(c)p(s)/p(c,s)] \\ &= \textstyle\sum_{(c,s)} p(c)p(s) \ln [p(c)p(s)/p(s)p(c|s)] \end{aligned} \quad (11a)$$

Using Eq. (4),

$$D[(C)(S)\|(C,S)] = -\textstyle\sum_i p(s_i) \sum_k p(c_k) \Delta E_m(c_k|s_i) \quad (11b)$$

and from Eq. (6),

$$D[(C)(S)\|(C,S)] = \textstyle\sum_i p(s_i) \langle \Delta E_m(C|s_i) \rangle \quad (11c)$$

Therefore, the directed divergence $D((C)(S)\|(C,S))$ is the mean potential of all conformations averaged over all possible sequences. We denote this divergence (Eq. 11c) as $\langle D_{\text{thread}}(S,C) \rangle$. This quantity is the average of the energies resulting from a series of threading exercises, in which many sequences are threaded through the conformation pool. We note that interest in the directed divergences between the two distributions (C,S) (the joint distribution) and $(C) \times (S)$ (their uncorrelated product) stems from the fact that

$$D[(C,S)\|(C)(S)] = \textstyle\sum_i p(s_i) D[(C|s_i)\|(C)] \quad (11d)$$

and

$$D[(C)(S)\|(C,S)] = \textstyle\sum_i p(s_i) D[(C)\|(C|s_i)] \quad (11e)$$

are simply sequence-averaged divergences of the two probability functions $p(c|s)$ and $p(c)$ for a specified sequence. The divergences $D[(C|s_i)\|(C)]$ and $D[(C)\|(C|s_i)]$, in measuring the distance between the sequence-independent and sequence-dependent distributions, quantify the effect of sequence on the discrimination of conformations.

The expected difference between $\Delta E(c^T|s_i)$ and $\langle \Delta E(C|s_i) \rangle$ is straightforward to calculate:

$$\begin{aligned} &\textstyle\sum_{s,c} \{\Delta E(c^T|s_i) - \langle \Delta E(C|s_i) \rangle\} p(s_i,c^T) \\ &\quad = \textstyle\sum_{s,c} \Delta E(c^T|s_i) p(s_i c^T) - \sum_{s,c} \langle \Delta E(C|s_i) \rangle p(s_i,c^T) = \\ &-I_g(S,C) - D_{\text{thread}}(S,C) = -D((S,C)\|(C)(S)) - D((C) \\ &\quad \times (S)\|(S,C)) = -J((S,C)\|(C)(S)) \end{aligned} \quad (12)$$

Given that the threading Z-score is defined as Eq. (7), the expected Z-score is related to the total divergence through the following relation:

$$\langle Z(S,C^T) \sigma(S,C) \rangle = -J((C,S)\|(C)(S)) = -I_g(C^T,S) - \langle D_{\text{thread}}(S,C) \rangle \quad (13a)$$

[derived by multiplying both sides of Eq. (7) by $\alpha(s_i,c)$ and then computing the expectation value of each side.] This equation links information gain, the quantity whose optimization we studied in previous work, to the Z-score, an established measure of threading performance.[18] We note here that the two variables in the bracket on the l.h.s. of Eq. (13a) above are uncorrelated, a statement we will prove in the Results section below. It follows that

$$\langle Z(S,C^T) \rangle \langle \sigma(S,C) \rangle = -J((C,S)\|(C)(S)) = -I_g(C^T,S) - \langle D_{\text{thread}}(S,C) \rangle \quad (13b)$$

The mean Z-score $\langle Z(S,C^T) \rangle$ has been identified as a gauge of the predictive power of a given force field.[29] We emphasize that the quantities involved in the above relation are *grand averages*, or averages over the entire sequence and structure domains. For instance, the quantity $\langle Z(S,C)$ $\alpha(S,C) \rangle$ is the expected product of Z and σ computed from a series of threading exercises using all the members of a pool of sequences, on a representative conformational pool.

METHODS

Demonstrating the Informatic Nature of Threading: Threading Short Sequence Segments

The measure of success of a scoring function in a threading exercise is its ability to assign the best score to the native conformation consistently. In previous work,[2] we demonstrated that the choice of sequence and structural descriptors affects the amount of information one can extract from a finite data set, by directly influencing the probability estimates used to calculate the associated entropies. The informatic relationship expressed in Eq. (13), summarized in Table I, and illustrated in Figure 1, extends this point to statistical potentials and threading. The way sequence and structure are parameterized affects the quantities $I_g(C,S)$, $\langle D_{\text{thread}}(S,C) \rangle$, $\langle J(S,C) \rangle$, $\langle \sigma(S,C) \rangle$, and $\langle Z(S,C) \rangle$, and will therefore influence the ability of the potential to select the native conformation.

One can therefore gauge how well a potential function performs by comparing the ranking of energies given by the native conformation and the pool of test structures. Because local effects account for only a fraction of the stabilization energy of a protein, potentials based solely on local sequence will not identify the native or true conformation with perfect accuracy. However, the relative rank of the energy assigned to the native conformation by a local potential should provide an indication of its effectiveness. We define this measure as $r_Z(s_i,c^T)$, the relative ranking of the Z-score of the native conformation c^T of sequence s_i in

TABLE I. Informatic and Energetic Quantities

Quantity	Instance	Mean
Energy of a particular conformation	$\Delta E(c_k \vert s) = -\ln p(c_k \vert s)/p(c_k)$	
Energy of a native interaction	$\Delta E(c_k{}^T \vert s) = -\ln p(c_k{}^T \vert s)/p(c_k{}^T)$	$\langle \Delta E(C^T \vert S)\rangle = \sum_{i,k} \Delta E(c_k{}^T \vert s_i)\, p(s_i,c_k{}^T)$
Energy of the native conformation	$\Delta E_m(c^{T*} \vert s) = (1/m) \sum_k^m \Delta E(c_k{}^T \vert s)$ where $c^{T*} = c_1{}^T c_2{}^T \ldots c_k \ldots c_m{}^T$	$\langle \Delta E_m(C^T \vert S_i)\rangle = \sum_{i,j} \Delta E_m(C^T \vert S)$ $\approx \langle \Delta E(C^T \vert S)\rangle$
Information gain (Divergence)		$I_g(S,C^T) = H(C^T) - H(C^T \vert S)$ $= \sum p(c^T) \ln p(c^T) - \sum p(c^T \vert s) \ln p(c^T \vert s)$ $= -\langle \Delta E(C^T \vert S)\rangle$
Threading divergence	$D_{\mathrm{thread}}(s,C) = \sum_{j\,U}^{n} \Delta E(c_j \vert s)\, p(c_j)$	$\langle D_{\mathrm{thread}}(S,C)\rangle = \sum_i D_{thread}(S_i\, C) p(s_i)$
Total divergence	$J(s,C^T) = D_{\mathrm{thread}}(s,C) - \Delta E(c^T \vert s)$	$\langle J(S,C)\rangle = \sum_i J(s_i,C^T)\, p(s_i)$ $= \langle D_{\mathrm{thread}}(S,C)\rangle + I_g(S,C^T)$
Spectral standard deviation	$\sigma(s,C)$	$\langle \sigma(s_i,C)\rangle = \sum_i \sigma(s_i,C)\, p(s_i)$
Z-score	$Z(s,c^T) = [\Delta E(c^T \vert s) - \sum_{j\,U}^{n} \Delta E(c_j \vert s)\, p(c_j)]/\sigma(s,C)$	$\langle Z(S,C^T)\rangle = \sum_i Z(s_i,c^T)\, p(s_i,c^T)$ $\approx [-I_g(S,C^T) - \langle D_{\mathrm{thread}}(S,C)\rangle]/\langle \sigma(s_i,C)\rangle$ $= -\langle J(S,C^T)\rangle/\langle \sigma(s_i,C)\rangle$
Relative ranking of Z-score (or energy)	$r_Z(s,c^T)$	$\langle r_Z(S,C^T)\rangle = \sum_i r_Z(s_i,c^T)\, p(s_i,c^T)$

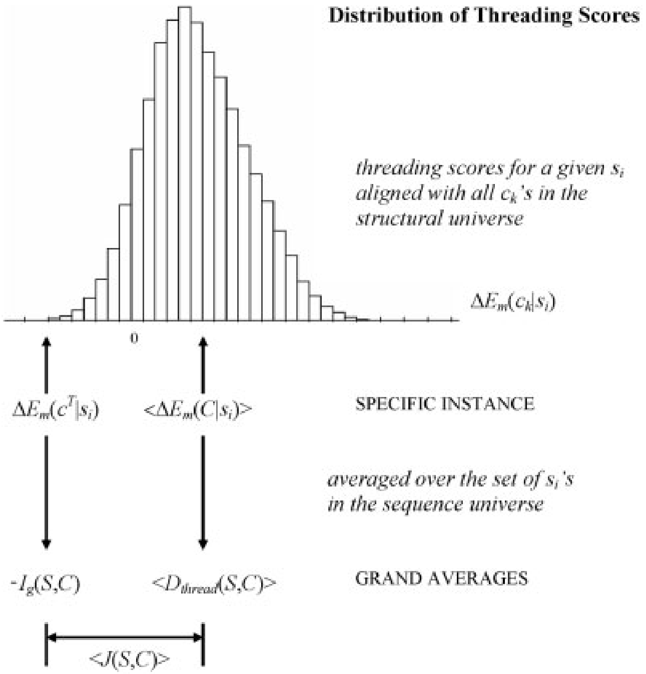
Distribution of Threading Scores
threading scores for a given s_i aligned with all c_k's in the structural universe
$\Delta E_m(c_k|s_i)$
0
$\Delta E_m(c^T|s_i)$
$<\Delta E_m(C|s_i)>$
SPECIFIC INSTANCE
averaged over the set of s_i's in the sequence universe
$-I_g(S,C)$
$<D_{thread}(S,C)>$
GRAND AVERAGES
$<J(S,C)>$

Fig. 1. The relationships among informatic quantities, statistical potentials, and the threading procedure. A typical spectrum of energy scores given by the ensemble of decoy conformations in the test set shows the relative position of the native energy $\Delta E(c^T|s_i)$ against the spectral mean $\langle \Delta E_m(S|s_i)\rangle$. Threading a series of sequences, one arrives at a set of these energy spectra, from which grand averages $I_g(C,S)$, $\langle D_{\mathrm{thread}}(S,C)\rangle$, $\langle J(S,C)\rangle$ can be obtained. In addition, every spectrum is characterized by the standard deviation $\sigma(s_i,C)$, each contributing to the average quantity $\langle \sigma(S,C)\rangle$. Consequently, the mean Z-score $\langle Z(S,C)\rangle$ can be calculated via Eq. (13).

the conformational pool. The quantity $r_Z(s_i,c^T)$ is also the energy ranking, as the equivalence between Z-score and energy [shown in Eq. (7)] involves only constants for a particular threading exercise, and should therefore not affect score ranking. {A value $r_Z(s_i,c^T) = 1$ means that the c^T is assigned the lowest energy and is therefore ranked 1; consequently, $\max[r_Z(s_i,c^T)] = n_U$, where n_U is the number of conformations in the test pool.} For each set of conditions, we will examine both the traditional gauge of threading success [i.e., how often $r_Z(s_i,c^T) = 1$] and the rank average, $\langle r_Z(S,C^T)\rangle$, given by a battery of threading exercises, as measures of the performance of statistical potentials.

We consider the threading of short sequences instead of whole protein chains, to ensure that there is sufficient diversity of conformations in the test pool. It is appropriate to measure the effectiveness of a local potential by using only short segments, because such potentials measure only local side-chain–backbone interactions.

Sequence and Structural Partitions Used in this Work

We examine the threading performance of energy functions derived from four levels of resolution of the Ramachandran space: the crudest partition $(\phi,\psi)_{60°}$, subdivides the space into a 6 × 6 field, bounded by 60° by 60° squares, with grids at $\{60n - 180, n = 0, \ldots, 6\}$. The other three are $(\phi,\psi)_{20°}$, $(\phi,\psi)_{10°}$, and $(\phi,\psi)_{7.5°}$, with grids at appropriate intervals (always including the axes $\phi = 0°$ and $\psi = 0°$ as grid boundaries) to subdivide the space into 18 × 18, 36 × 36, and 48 × 48 fields, respectively.

To specify local sequence, we use three different descriptors. The simplest is the single-residue sequence, which gives only the identity of each residue. We refer to this sequence descriptor as S_1. The other two categories specify the local trimer sequence at different resolutions. One uses the complete 20 amino acid alphabet for each position in the trimer, and the other uses optimized contractions of the amino acid alphabet at the two ends of the trimer segment, and the 20 amino acid alphabet for the middle residue. We refer to these as $S_{20\text{-}20\text{-}20}$ and $S_{x\text{-}20\text{-}y}$, respectively with x and y indicating the number of amino acid

TABLE II. Information Gain for Different Structural and Local Sequence Partitions

Local sequence partition[b]	Structural Partition/Resolution[a]			
	60°	20°	10°	7.5°
S_1	0.1591 nats	0.2013	0.1618	0.1246
$S_{x\text{-}20\text{-}y}$	0.2098	0.2612	0.2358	0.2186
$S_{20\text{-}20\text{-}20}$	0.1828	0.2369	0.2045	0.1766

[a] Refers to the size of the square bins on each side that subdivides the (ϕ,ψ) space.

[b] The subscripts of the sequence desciptor S describe the extent of local sequence description. Subscript 1 refers to single-residue description; x-20-y is the trimer sequence description, with contracted amino acid alphabet at the flanking positions; 20-20-20 is the full amino acid sequence description at all three trimer position.

TABLE III. The Most Informative Contracted Amino Acid Alphabets for Trimer Sequence Configuration

Structural resolution	Sequence partition $S_{x\text{-}20\text{-}y}$	Amino acid membership
60°	$x = 3$[a]	1 A C D E H K V W Y
		2 F I L M N P Q R S T
		3 G
	$y = 7$[a]	1 D E H K N Q R S
		2 C F L M W Y
		3 I V
		4 G
		5 P
		6 A
		7 T
20°	$x = 3$	1 A D E H K N P Q R S T
		2 C F I L M V W Y
		3 G
	$y = 7$	1 F I L V W Y
		2 D H N S T
		3 E K Q R
		4 A M
		5 G
		6 P
		7 C
10°	$x = 3$	1 A D E H K N P Q R S T
		2 C F I L M V W Y
		3 G
	$y = 10$	1 C F I L V W Y
		2 E K Q R
		3 A M
		4 G
		5 P
		6 T
		7 D
		8 N
		9 H
		10 S

[a] These values of x and y are the numbers of amino acid groups, at the amino and carboxyl ends, of the local trimer sequence that give the most information gain. The middle position of the trimer is described by the full 20 amino acid alphabet. In the text, this partition is notated as $S_{x\text{-}20\text{-}y}$.

clusters at the amine and carboxyl end of the trimer, respectively. The latter representation makes use of the information maximization algorithm, which we developed in previous work[3] to cluster the amino acids.

Generating Probability Distributions for Use in Statistical Potentials

The probability distributions arising from the different combinations of sequence and structural partitions described above provide the basis for the local backbone potential function. The probability distributions generated are similar to those pictured in Figure 8 of ref. 3. To form unbiased potentials, the protein chain to be analyzed is first taken out of the data set before the probability distributions are computed from the remaining proteins.

For each of the four (ϕ, ψ) resolutions, the sequence-independent probability distribution is constructed, together with the set of local sequence-dependent probability distributions, from the protein data set. We refer the reader to previous work[3] in which the algorithm used to build these distributions is described. One can compute the backbone energy associated with any sequence-structure combination from these probability distributions, using Eq. (5).

RESULTS AND DISCUSSION

I. Information Gain

Comparison of Information Gain Among Partitions

The information gains given by the three resolutions (60°, 20°, and 10°), calculated from Eq. (3c) for the three different types of local sequence description, are summarized in Table II. These values illustrate the competing constraints in maximizing the information that can be extracted from finite data sets. Overpartitioning of either the sequence or structural domain degrades information because the need for greater detail in description is countered by a need for more data to generate meaningful statistics. Underpartitioning will neglect detail that provides relevant information. One finds optimal conditions between these partitioning extremes.[2] Examining the effect of local sequence on backbone conformation, we find that the current data set yields the most amount of information using (1) the 20° phi–psi angle partition, and (2) the trimer sequence level with alphabet contraction in the outer positions ($S_{x\text{-}20\text{-}y}$). In further analyses, we examine how this informatically optimal scheme compares with other sequence and structure partitions in assigning low (favorable) energies to the native conformation, and in accurate detection of the native state in a threading exercise.

Amino Acid Alphabet Contraction for $S_{x\text{-}20\text{-}y}$

The amino acid groupings that yield the highest information gain in configuration $S_{x\text{-}20\text{-}y}$, tested at three resolution levels, are summarized in Table III. The Monte Carlo search procedure that generated these data explored different values of x and y (ranging from 2 to 15 clusters at each position), as well as different amino acid groupings within each given x and y. Consistent with previous observations,[3] the residue at the carboxyl end of the trimer exerts more influence on the phi-psi conformation of the middle residue, and therefore using a larger number of clusters at this position increases the effective information gain. In all three structural partitions, we observe that maximal information is obtained when $x < y$.

The details of these amino acid clusterings are also consistent with previous work.[2] We summarize our results as follows:

1. Although the overall amino acid grouping at the amino terminal end of the trimer for the 60° resolution is different from those at 20° and 10°, tight groupings {ADEHK}, {NPQRST}, {FILM}, and {VWY} are consistent at all three structural levels. It is possible that these groupings will be more clearly expressed as more data become available, when an increase in x (the number of groupings in the amino-terminal position) can be accommodated.
2. The unique properties of glycine force the optimization procedure to consistently recognize it as a separate cluster at both ends of the trimer. Similarly, the considerable influence of proline at the carboxyl end of the trimer is also expressed by the clustering results.
3. The fundamental division between polar and hydrophobic amino acids is maintained, as exemplified by the tight groups {EKQR} and {FILVWY}.
4. The information maximization algorithm does not cleanly partition the amino acids into reproducible groupings at the structural levels investigated. Although similarities in local coding exist among amino acids, it is safe to assume that each amino acid bears enough individual traits to make it informatically unique as far as its effect on backbone conformation is concerned. As more data become available, the information maximization algorithm should be able to accommodate a larger number of clusters (i.e., larger values of x and y), which should allow the finer local coding properties of each amino acid to be recognized.

II. Information Gain and Statistical Potentials

We illustrate the effect of optimizing $I_g(S,C)$ on local backbone energy. The resolution levels $(\phi,\psi)_{60°}$, $(\phi,\psi)_{20°}$, and $(\phi,\psi)_{10°}$ are again used to illustrate the effect, while the sequence component of the conditional probability $P(C|S)$ is represented as $S_{x\text{-}20\text{-}y}$ (the alphabet-contracted trimer configuration). From the information gains (Table II), the average net energies $\langle \Delta E(\mathrm{C}_{(\phi,\psi)}|S_{x\text{-}20\text{-}y})\rangle$ due to the trimer amino acid sequence are -0.2098, -0.2612, and $-0.2358kT$, respectively [at the corresponding (x,y) pairs (3,7), (3,7), and (3,10)]. The desirability of assigning low energy values to the native conformation suggests that the 20° resolution may be the most effective bin level to employ as a parameter for statistical potentials.

To examine how the choice of structural description affects the assignment of energy values to native conformation, we undertake two calculations: (1) the total backbone energy of whole protein chains, and (2) the total backbone energy of chain segments of length 10.

1. *Trimer-sequence-dependent backbone energy of whole protein chains.*

Backbone torsional energy was calculated for the native conformation of each chain in the data set, using the optimal probability distributions derived at the three resolutions. To illustrate the distribution, the histograms of energies are shown in Figure 2. The mean backbone energy per position using partitions 60°, 20°, and 10° are -0.2142, -0.2674, and $-0.2441kT$, quite close to the grand averages $\langle \Delta E(C_{(\phi,\psi)}|S_{x\text{-}20\text{-}y})\rangle$ listed above. Although most of the proteins have been assigned negative values for the backbone energy, 29, 27, and 43 proteins, respectively, show positive $\langle \Delta E(C_{(\phi,\psi)}|S_{x\text{-}20\text{-}y})\rangle$. Clearly, the occurrence of frustration in the choice of local conformation (as evidenced by positive values of the backbone energies) is compensated for by other types of interactions in order for the proteins to have negative total energies. However, selecting the 20° resolution (vs. 60° or 10°) reduces the number of proteins assigned unfavorable backbone energies on a purely local basis.

Next, we rank the backbone energies assigned to each protein chain by the three (ϕ,ψ) resolutions. We find that out of a total of 1045 chains in the data set, the partition $(\phi,\psi)_{20°}$ assigns the lowest energy to 716 chains (68.5%). The rest of the chains are assigned lowest energy values by the other two partitions. The partitions $(\phi,\psi)_{60°}$ and $(\phi,\psi)_{10°}$ produce lowest values for 127 (12.2%) and 202 chains (19.3%).

2. *Trimer-sequence-dependent backbone energy of chain segments of length* 10.

The behavior of the backbone energy of the native conformation of 10-mer chain segments provides further insight into the performance of the optimized probability distributions. We calculate the specific information gain (or "energy") $\Delta E_m(c^T|s_{x\text{-}20\text{-}y})$ for the *native* conformation C^T of a sequence segment of length 12 with a conformation of length $m = 10$ (because a sequence segment of 12 residues is necessary to thread 10 overlapping trimer segments, and therefore 10 trimer-(ϕ,ψ) pairs). We expect the average net energy of 10-mers $\langle \Delta E_{10}(C^T|S_{x\text{-}20\text{-}y})\rangle$ to be close to the overall average $\langle \Delta E(C^T|S_{x\text{-}20\text{-}y})\rangle$, but individual values of $\Delta E_{10}(c^T|s_{x\text{-}20\text{-}y})$ will scatter around the mean, some with positive energies. Out of a total of 187,804 10-mers in the data set, the lowest energy is assigned to 78,301 10-mers by $(\phi,\psi)_{20°}$ (41.7%), while 49,495 (26.4%) and 60,008 (31.9%) 10-mers have lowest backbone energies using $(\phi,\psi)_{60°}$ and $(\phi,\psi)_{10°}$, respectively. Details of the comparison are outlined in Table IV.

III. Information Gain and Threading

One might argue that any change in statistical potential parameters that brings about an increase in native structure probability implies a concomitant decrease in the probabilities of nonnative conformations. The selectivity of the energy function can only be determined by examining its performance against the spectrum of energy values assigned to alternative (nonnative) conformations. We investigate this issue by examining the mechanics of threading short sequence segments.

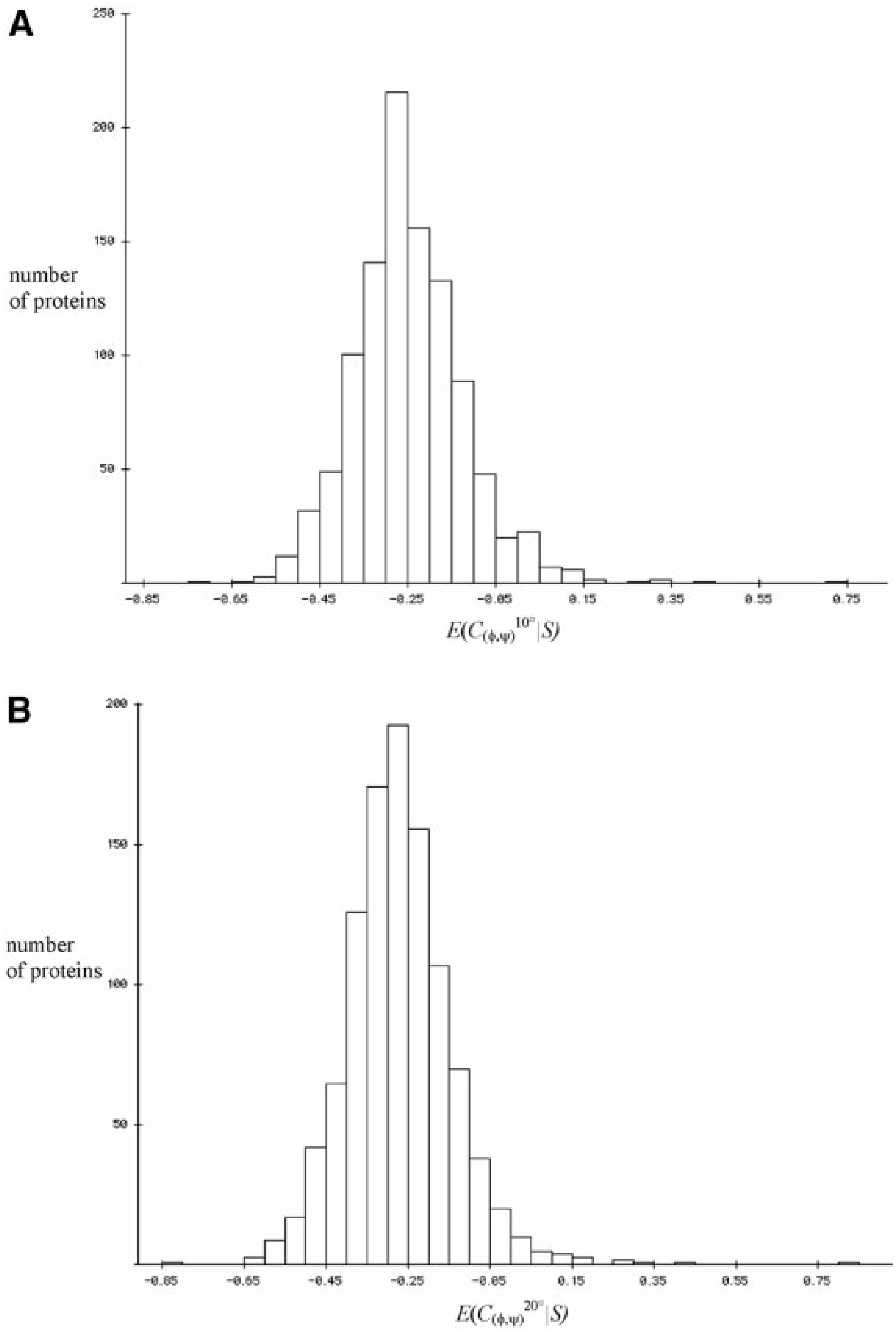

Fig. 2. Average backbone energy of whole protein chains. Trimer sequence-dependent backbone torsional energy was calculated for the native conformation of every chain in the data set at the three resolutions (**A**) 10°, (**B**) 20°, and (**C**) 60°. Although most of the proteins have been assigned negative values for the backbone energy, 43, 29, and 27 proteins, respectively, show postive $\langle \Delta E(C_{(\phi,\psi)}|S_{x\text{-}20\text{-}y})\rangle$. Selecting the 20° resolution (vs. 60° and 10°) reduces the number of proteins assigned unfavorable backbone energies on a purely local basis.

The Threading Procedure

A threading calculation for short chain segments proceeds by pairing every conformation c_k of length 10 in the test pool with the query 12-mer amino acid sequence s_i (noting that the first and last amino acid residues are necessary to specify the trimer sequence configuration of the terminal conformations). When energy quantities measuring the fit between the query sequence and the test conformation are computed for each conformation in the pool, the average value $\langle \Delta E_{10}(C_k|s)\rangle$ [or threading divergence $D_{\text{thread}}(s,C)$] and the standard deviation $\sigma(s_i,C)$ of the distribution are obtained straightforwardly. The Z-score follows easily via Eq. (7).

A sample calculation is done for amino acid segment 163–174 of the A chain of the protein 1c8e (feline panleuko-

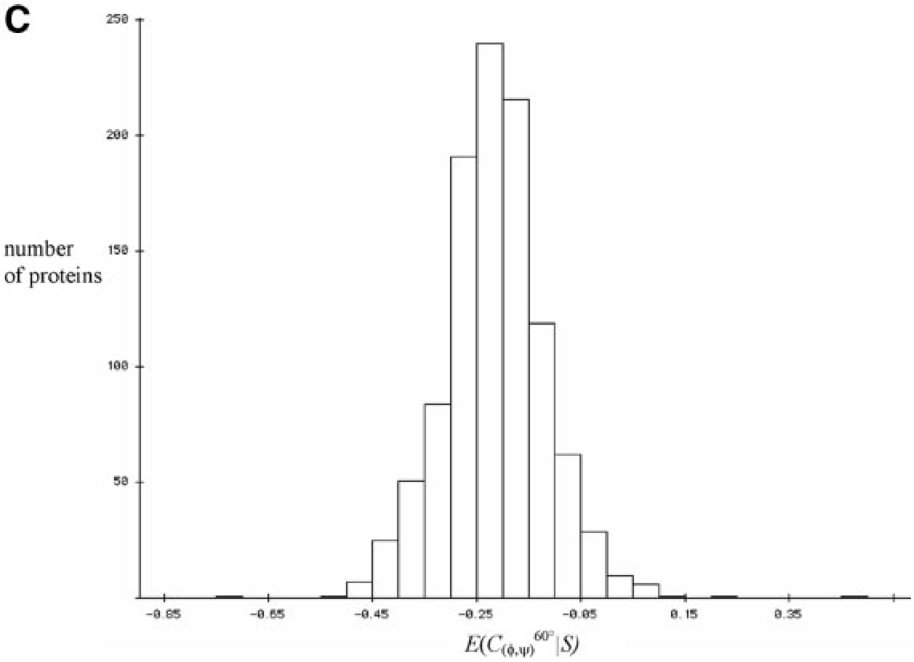

Figure 2. (Continued.)

TABLE IV. Comparing Energy Values of 10-mers Using Three Structural Partitions

Number of 10-mers assigned:	Structural partition/resolution		
	60°	20°	10°
The best energy	49,495	78,301	60,008
The 2nd best energy	39,434	75,902	72,468
The worst energy	99,043	33,517	55,244
$\langle \Delta E_{10}(C^T\|S)\rangle$	−0.2142	−0.2674	−0.2441
$\langle Z(S,C^T)\rangle$	−1.494	−1.827	−1.890
$\langle r_Z(S,C^T)\rangle$[a]	20704	16230	18115
$\langle \sigma(s_i,C)\rangle$	0.3367	0.3249	0.2671

[a] Out of a total of 187,804 decoy structures in the test set.

penia virus empty capsid structure). The amino acid sequence and (ϕ,ψ) conformation are shown in Table V. The energies of the native structure are -0.694, -0.601, -0.564, and $-0.576kT$, respectively for the resolutions $(\phi,\psi)_{60^\circ}$, $(\phi,\psi)_{20^\circ}$, $(\phi,\psi)_{10^\circ}$, and $(\phi,\psi)_{7.5^\circ}$.

The amino acid sequence is threaded through every 10-mer conformation c_k in the data set, giving a total of 187,804 trials. Each (s_i,c_k) pairing results in an energy score $\Delta E_m(c_k|s_i)$. The distribution of these scores for the partition $(\phi,\psi)_{20^\circ}$, shown in Figure 5(B), is characterized by the average $\langle \Delta E_{10}(C|s_i)\rangle = D_{\text{thread}}(s,C) = 0.656$ and standard deviation $\sigma(s_i,C) = 0.425$. From Eq. (7), the Z-score for C^T is -2.958. We also rank all the alternative conformations with respect to their Z-scores, and find that only 64 conformations have better energy values (i.e., $r_Z(s,c^T) = 64$). Similar calculations were performed for the other structural partitions, with different threading outcomes. The results for this segment, threaded using the $(\phi,\psi)_{20^\circ}$ partition, are summarized in Table V.

General Threading Results

The foregoing outlines the procedure for threading a particular 10-mer sequence through the pool of conformations. We now examine threading patterns by performing a battery of threadings of an ensemble of sequence segments, to compute the grand averages of the informatic quantities in the third column of Table I. The best estimate for these averages is obtained when all possible sequence segments (totalling 187,804 in our data set) are threaded through the conformation pool. This is how we compute averages in this work.

We compare the overall performance of 10-mer threading using empirical potential functions derived from the following four structural partitions: $(\phi,\psi)_{60^\circ}$, $(\phi,\psi)_{20^\circ}$, $(\phi,\psi)_{10^\circ}$, and $(\phi,\psi)_{7.5^\circ}$, recalling that $(\phi,\psi)_{20^\circ}$ gives the best $I_g(S,C^T)$ among the four. From the equivalence between $I_g(S,C^T)$ and $\langle \Delta E(C^T|S)\rangle$, already established, we observe that $(\phi,\psi)_{20^\circ}$ also exhibits the best mean energy assigned to the native conformation. These results are summarized in Table VI. Comparison of the mean ranking of the native conformation score against the universe of structures in the full data set reveals that the informatically superior partition $(\phi,\psi)_{20^\circ,}$ on average, performs best in assigning low energy to the correct conformation among the pool of conformations in the data set. Moreover, the relative ranking of success measured by $\langle r_Z(S,C^T)\rangle$ mirrors the ranking of $\langle \Delta E(C^T|S)\rangle$. The Z-score values were also averaged to assess their predictive value with respect to threading success. Interestingly, the average Z-score does

TABLE V. Threading Results for 10-mer Segment 163–172 of Chain 1c8eA

	162	163	164	165	166	167	168	169	170	171	172	173
Amino acid Sequence	Y	P	W	K	P	T	I	P	T	P	W	R
Secondary Structure[a]	C	T	T	S	C	B	C	C	E	E	E	E
$\phi°$	−120	−55	−92	−127	−48	−129	−163	−88	−106	−57	−90	−154
$\psi°$	125	−39	20	110	148	−159	157	167	146	130	136	154

	$S_{x\text{-}20\text{-}y} - (\phi,\psi)_{10}*$	$S_{x\text{-}20\text{-}y} - (\phi,\psi)_{20}*$	$S_1 - (\phi,\psi)_{20}*$
$\Delta E_{10}(c^{T*}\|s)$	−0.564	−0.601	−0.546
$D_{\text{thread}}(s,C)$	0.436	0.656	0.935
$\sigma(s,C)$	0.296	0.425	0.662
$Z(s,c^T)$	−3.378	−2.958	−2.237
$r_z(s,c^T)$	89	64	103
$J(s,C^T)$	1.000	1.257	1.481

[a] DSSP description, as described in ref. 30.

TABLE VI. Summary of Threading Results of 187,804 10-mer Sequences Using Different Sequence and Structural Partitions

Local sequence	(ϕ,ψ) resolution	$\langle\Delta E(C^T\|S)\rangle$	$\langle\sigma(s_i,C)\rangle$	$\langle r_z(S,C^T)\rangle$	$\langle Z(S,C^T)\rangle$	$\langle D_{\text{thread}}(S,C)\rangle$	$\langle J(\text{S,C})\rangle$	$n(r_z(s,c^T) = 1)$	$n(r_z(s,c^T) \leq 10)$
$S_{x\text{-}20\text{-}y}$	60°	−0.214	0.337	20704	−1.494	0.289	0.503	891	4818
$S_{x\text{-}20\text{-}y}$	20°	−0.267	0.325	16231	−1.827	0.326	0.594	2140	9705
$S_{x\text{-}20\text{-}y}$	10°	−0.244	0.272	18115	−1.890	0.269	0.513	1929	8961
$S_{x\text{-}20\text{-}y}$	7.5°	−0.228	0.249	19697	−1.866	0.236	0.465	1693	8231
$S_{20\text{-}20\text{-}20}$	60°	−0.186	0.237	22040	−1.584	0.190	0.376	857	4379
$S_{20\text{-}20\text{-}20}$	20°	−0.241	0.256	16921	−1.910	0.247	0.489	1726	8187
$S_{20\text{-}20\text{-}20}$	10°	−0.212	0.241	19933	−1.898	0.246	0.458	1114	6414
$S_{20\text{-}20\text{-}20}$	7.5°	−0.185	0.233	22344	−1.842	0.243	0.429	808	4907
S_1	60°	−0.162	0.290	23643	−1.320	0.221	0.383	732	3778
S_1	20°	−0.207	0.406	18293	−1.400	0.361	0.568	1833	8162
S_1	10°	−0.172	0.466	22326	−1.312	0.439	0.611	1877	8084
S_1	7.5°	−0.137	0.497	25441	−1.250	0.484	0.621	1637	7514

not follow the pattern set by $I_g(S,C^T)$. The best average Z-score occurs at the $(\phi,\psi)_{10°}$ resolution, even though the $(\phi,\psi)_{20°}$ resolution performs better in threading, as measured by $\langle r_Z(S,C^T)\rangle$ (Table VI).

The expression in Eq. (13a) relates the average product of the Z-score and the associated standard deviation of the spectrum to the total divergence. The cumbersome left-hand side can only be deconvoluted into a product of averages if the two quantities [Z-score and $\sigma(s_i,C)$] are uncorrelated. Pairs of quantities from 9445 threading applications using sequence and structural descriptors $S_{x\text{-}20\text{-}y}$ and $(\phi,\psi)_{20°}$ are plotted in Figure 3. The Pearson product-moment correlation, which equals −0.071, confirms that the two quantities are indeed uncorrelated. Direct verification by actual calculation of the three quantities $\langle Z(S,C^T)\rangle$, $\langle\sigma(s_i,C)\rangle$, and $\langle Z(S,C^T)\ \sigma(s_i,C)\rangle$ across all 187,804 threading applications (using the same sequence and structural desciptors) produces the values −1.8270, 0.3249, and −0.5961, respectively, from which we calculate the product $\langle Z(S,C^T)\rangle\ \langle\sigma(s_i,C)\rangle = -0.5936$, closely approximating the value $\langle Z(S,C^T)\ \sigma(s_i,C)\rangle = -0.5961$. This equivalence is observed consistently in every potential function examined in this work. Therefore, it is safe to assert that Eq. (13b) holds.

We extend the analysis of threading behavior to the two other types of sequence partitions: S_1 and $S_{20\text{-}20\text{-}20}$. Results summarized in Table VI show that ranking in $\langle\Delta E(C^T|S)\rangle$ is perfectly predictive of the performance as measured by $\langle r_Z(S,C^T)\rangle$, while $\langle Z(S,C^T)\rangle$ is not, even though the structural partition $(\phi,\psi)_{20°}$ shows the best $I_g(S,C)$, $\langle\Delta E(C^T|S)\rangle$, and $\langle r_Z(S,C^T)\rangle$ in both sequence partitions.

Behavior of Informatic Quantities in Threading

We test the effectiveness of the local potentials in assigning the lowest energy to the native conformation, bearing in mind that such potentials are incomplete, due to the absence of any long-range sequence information. We consider two quantities: $n(r_Z(s,c^T) = 1)$ and $n(r_Z(s,c^T) = 10)$, which refer to, respectively, the number of 10-mer sequences in which the rank of the native conformation

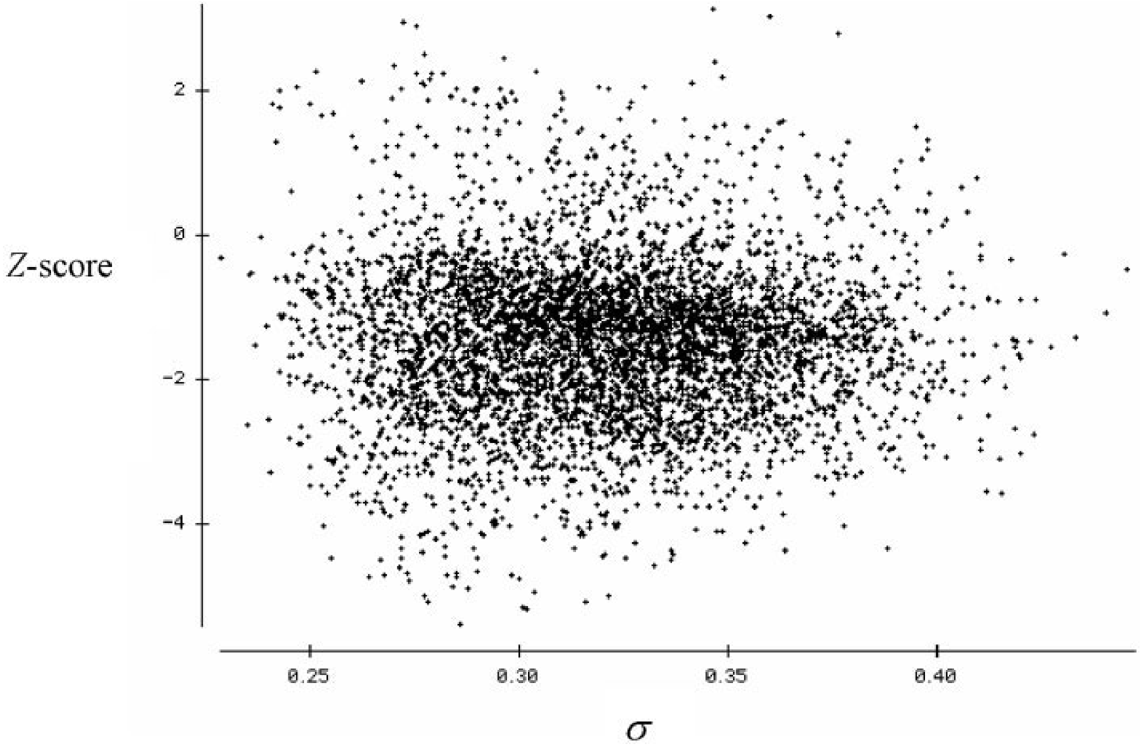

Fig. 3. *Z*-score and the threading spectral standard deviation. The plot shows the *Z*-score and $\sigma(s_i,C)$ for 9445 threading applications using sequence and structural descriptors $S_{x\text{-}20\text{-}y}$ and $(\phi,\psi)_{20°}$. The Pearson product-moment correlation of −0.071, confirms that the two quantities *Z*-score and $\sigma(s_i,C)$ are indeed uncorrelated, confirming the approximation used in Eq. (13b).

energy is 1 (best overall energy value); and the number of 10-mer sequences whose native conformational energy is among the 10 best, out of a total 187,804 different 10-mer sequences in the entire data set. These numbers are included in Table VI.

The potential that matches the greatest proportion of 10-mer sequences to their native structures [i.e., $n(r_Z(s,c^T) = 1)$] uses the $S_{x\text{-}20\text{-}y}$ sequence description and $(\phi,\psi)_{20°}$ resolution, with 2140 10-mers (or 1.14% of the total 10-mers in the data set). Likewise, the same potential assigns the most number of 10-mers one of the ten best energy values, at $n(r_Z(s,c^T) = 10) = 9705$ (or 5.17%). We note that this potential shows the lowest $\langle\Delta E(C^T|S)\rangle$ [and also the highest $I_g(S,C^T)$] among all potentials tested. We also calculate the correlation between these two measures and various informatic quantities, using the Spearman rank correlation, a nonparametric measure. The results, summarized in Table VII, reveal that $\langle\Delta E(C^T|S)\rangle$, $\langle D_{\text{thread}}(S,C)\rangle$, and $\langle J(S,C)\rangle$ show some measure of correlation with $n(r_Z(s,c^T) = 1)$ and $n(r_Z(s,c^T) = 10)$.

Next, we examine the average rank of the native conformation energy given by all 187,804 10-mers. Plots of the four informatic quantities $\langle\Delta E(C^T|S)\rangle$, $\langle D_{\text{thread}}(S,C)\rangle$, $\langle J(S,C)\rangle$, and $\langle Z(S,C^T)\rangle$ against the threading performance measure $\langle r_Z(S,C^T)\rangle$ for all kinds of sequence and structural partitions are summarized in Figure 4(A–D). Among these quantities, only $\langle\Delta E(C^T|S)\rangle$ is consistently predictive of $\langle r_Z(S,C^T)\rangle$, an observation confirmed by comparing the correlations between $\langle r_Z(S,C^T)\rangle$ and the various informatic quantities. This suggests that the sequence-structure partition that assigns the lowest average energies to the native conformation performs best in a threading procedure. Because of the equivalence between $\langle\Delta E(C^T|S)\rangle$ and $I_g(S,C^T)$, efforts to optimize information extraction, by maximization of information gain, should, in principle, increase the effectiveness of the potential function in a threading procedure.

TABLE VII. Spearman Rank Correlation between Native Conformation Energy Rank and Various Informatic Quantities

	$\langle r_z(S,C^T)\rangle$	$n(r_z(s,c^T) = 1)$	$n(r_z(s,c^T) \leq 10)$
$\langle\Delta E(C^T\|S)\rangle$	0.937	−0.580	−0.700
$\langle D_{\text{thread}}(S,C)\rangle$	−0.077	0.622	0.420
$\langle J(S,C)\rangle$	−0.168	0.741	0.580
$\langle Z(S,C^T)\rangle$	0.650	−0.140	−0.378
$\langle\sigma(s_i,C)\rangle$	0.112	0.427	0.182

On closer inspection, $\langle J(S,C)\rangle$ seems inversely proportional to $\langle r_Z(S,C^T)\rangle$, which is to be expected, except for two data points corresponding to the $[S_1-(\phi,\psi)_{10°}]$ and $[S_1-(\phi,\psi)_{7.5°}]$ sequence-structure partition pairs. For these two cases, the increase in $\langle J(S,C)\rangle$ is counteracted by a marked decrease in $\langle\Delta E(C^T|S)\rangle$, which is more significant in determining the ranking of the native conformation in the universe of structures. Last, the distributional position $\langle D_{\text{thread}}(S,C)\rangle$ shows little correlation with $\langle r_Z(S,C^T)\rangle$.

The plot of $\langle Z(S,C^T)\rangle$ versus $\langle r_Z(S,C^T)\rangle$ [Fig. 4(D)] shows wide variation in $\langle r_Z(S,C^T)\rangle$ at low values of $\langle Z(S,C^T)\rangle$,

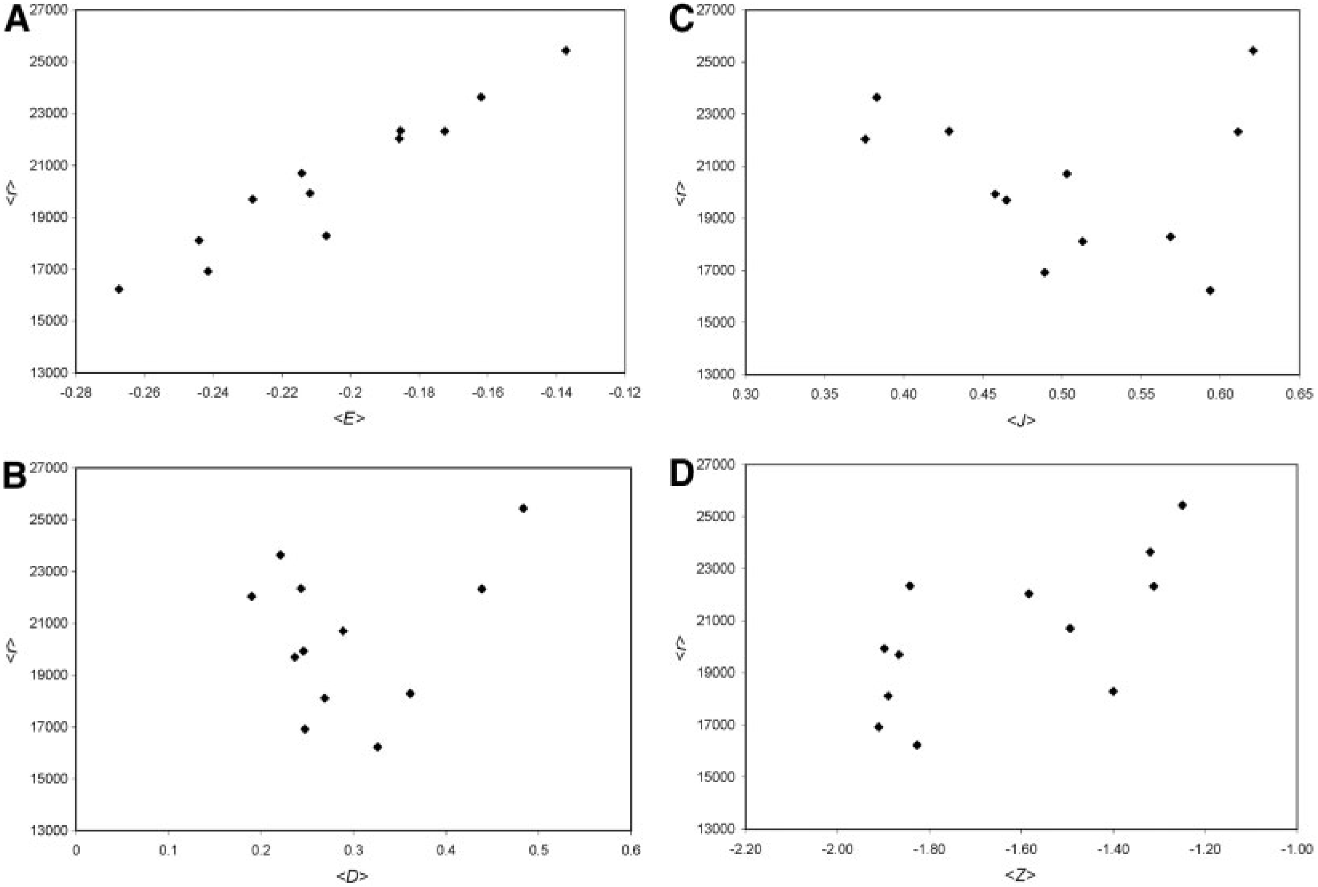

Fig. 4. Performance of four informatic quantities in predicting threading success. The quantities are plotted against the threading performance measure $\langle r_Z(S,C^T)\rangle$ for all sequence and structural partitions studied in this work: (**A**) $\langle \Delta E(C^T|S)\rangle$, (**B**) $\langle D_{\text{thread}}(S,C)\rangle$, (**C**) $\langle J(S,C)\rangle$, and (*D*) the standard mean of the *Z*-score $\langle Z(S,C^T)\rangle$. Among them, only $\langle \Delta E(C^T|S)\rangle$ is consistently predictive of $\langle r_Z(S,C^T)\rangle$, showing that the configuration that assigns low energies to the native conformation performs best in a threading procedure.

indicating that minimizing $\langle Z(S,C^T)\rangle$ does not necessarily decrease $\langle r_Z(S,C^T)\rangle$. This suggests that measuring the predictive power of force fields using the average *Z*-score[29] may not be appropriate in all instances. A number of studies have used the harmonic mean of the *Z*-score as an objective function by which to measure threading success,[18,30] defined as

$$\langle Z_{harm}(S,C^T)\rangle = n \;/\; \sum_{i}^{n} \,[1/Z(s_i,c^T)] \qquad (14)$$

Using this alternative definition does not make the *Z*-score more predictive of threading success, showing even less correlation with $\langle r_Z(S,C^T)\rangle$ than in Figure 4(D) (data not shown).

A Specific Example: Segment 163–172 of Chain 1c8eA

Based on threading success measures ($\langle r_Z(S,C^T)\rangle$, $n(r_Z(s,c^T)) = 1$, and $n(r_Z(s,c^T) = 10)$), we have shown that $\langle \Delta E(C^T|S)\rangle$ is especially predictive among the various informatic quantities available. To understand why this is so, we follow the threading behavior of one 10-mer segment, residues 163–172 of the protein 1c8eA, using three structural and sequence partitions. The energy values of the native conformation are -0.564, -0.601, and $-0.546kT$ for the descriptor pairs: $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10}$, $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$, and S_1–$(\phi,\psi)_{20°}$. The spectra of energy values assigned to alternative conformations in the universe of structures, for each of the three partitions, are illustrated in Figure 5. The standard deviations increase markedly (0.296, 0.425, and 0.662, in the order of the partitions enumerated above), and reflect the decreasing *Z*-score values computed for the native structures: -3.376, -2.954, and -2.238, despite the energy minimum at the $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ partition.

The relative ranking of the native conformation reveals that threading success follows the trend exhibited by energy and not *Z*-score. The partition $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ gives the lowest $r_Z(s,c^T)$, 64, while the other two partitions $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10°}$ and S_1–$(\phi,\psi)_{20°}$ give values of 89 and 103, respectively. A pictorial representation of the energy distribution given by the two partitions $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10°}$ and $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ (Fig. 6) shows that the bulk of the test conformations in the decoy pool lie to the right of the iso-energetic diagonal, indicating a general increase of energy values when one moves from the former to the latter partition. If the energy value of a given conformation

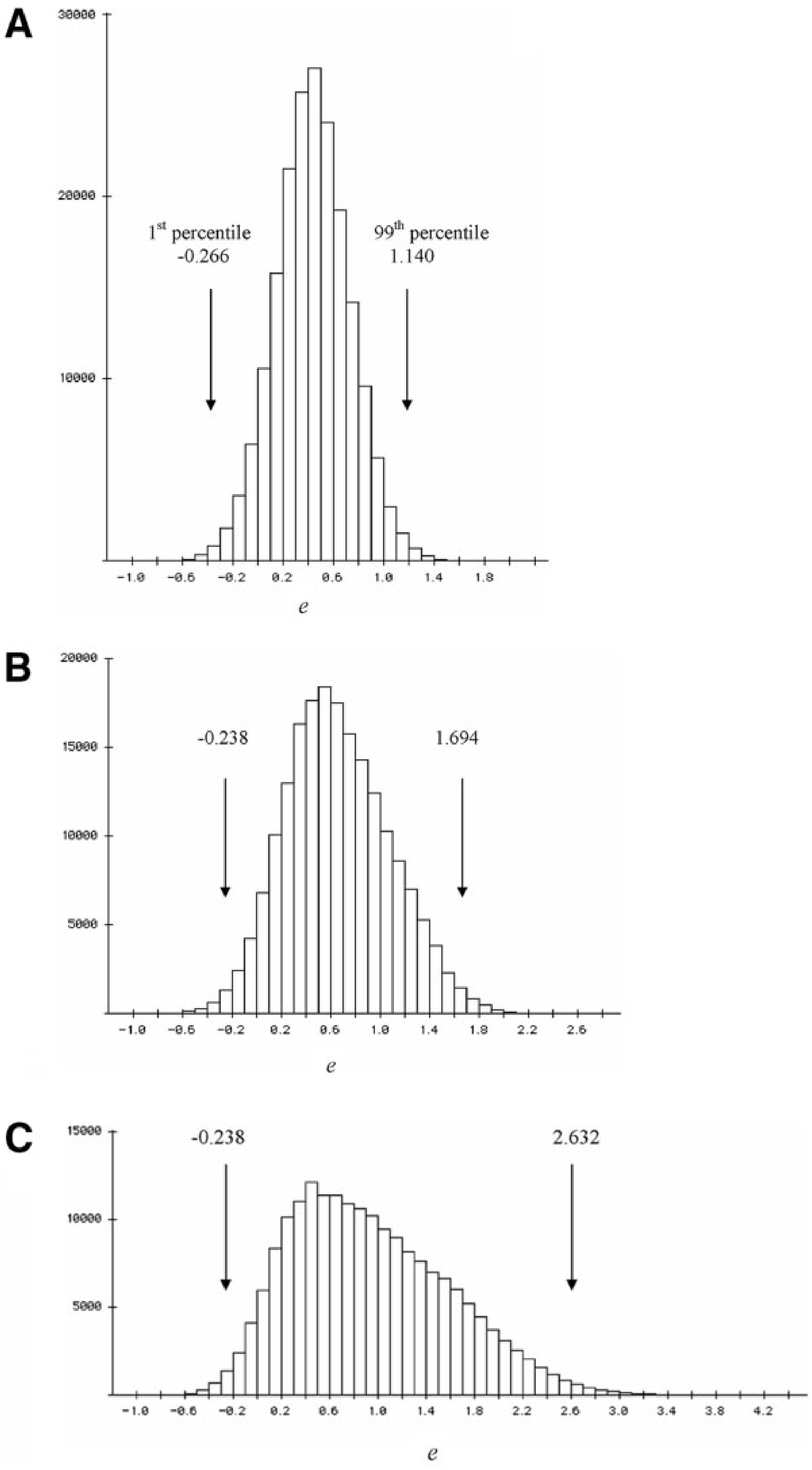

Fig. 5. Threading segment 163–172 of protein chain lc8eA. The spectra of energy values assigned to decoy conformations in the test pool, using three structural and sequence partitions: (**A**) $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10°}$, (**B**) $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$, (**C**) S_1–$(\phi,\psi)_{20°}$. The standard deviations increase markedly (0.296, 0.425, and 0.662, in the order of the partitions enumerated above), and reflect the decreasing *Z*-score values computed for the native structures: −3.376, −2.954, and −2.238, despite the energy minimum at the $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ partition. The first and 99th percentile values are indicated, showing that the widening of the distribution happens exclusively by stretching at the positive extreme, while keeping the negative ends anchored.

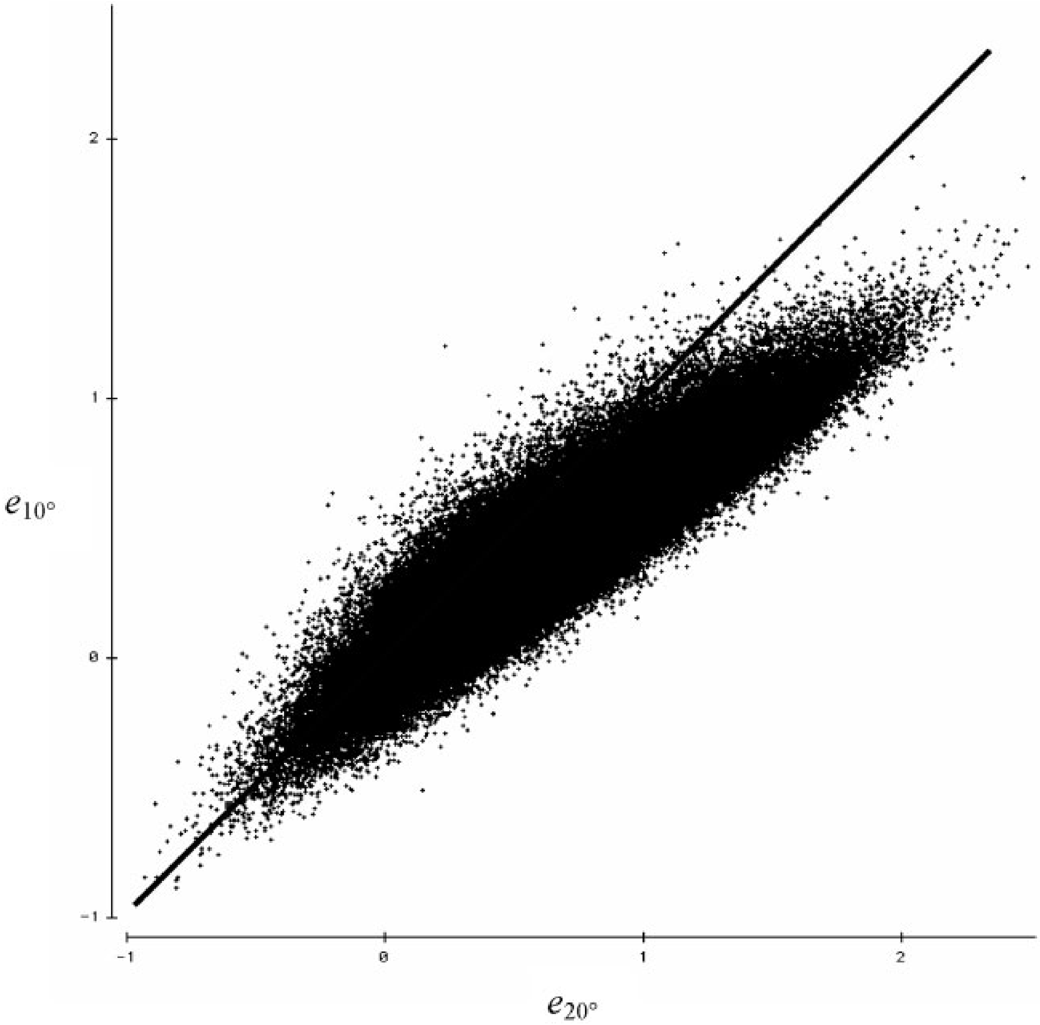

Fig. 6. The energy values assigned to decoy conformations in the tests pool, given by the two partitions $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10°}$ and $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$. The bulk of the test conformations in the decoy pool lie to the right of the iso-energetic diagnonal, indicating a general increase of energy values when one moves from the former to the latter partition.

is the same in both partitions (i.e., on the iso-energetic diagonal), the number of decoys having lower energy is greater in the former partition than the latter. This benefit is enhanced further by the accompanied lower energy value of the native conformation at $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ (-0.601) compared to the value at $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{10°}$ (-0.564).

Comparing the threading behavior at these two partitions, we observe a general increase in energy values for those incorrect conformations, shifting the distribution towards the positive (unfavorable) extreme (shown in Fig. 5). Assigning more unfavorable energies to the bulk of the conformational space is embodied quantitatively in the increase in standard deviation. What is beneficial to discrimination of the native structure, in this case, is not reflected in the behavior of the Z-score measure, which decreases in absolute value as a response to an increase in distribution width. This example illustrates that Z-score may be an inconsistent predictor of threading performance.

On the other hand, in comparing the threading behavior at partitions $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ and S_1–$(\phi,\psi)_{20°}$, we observe that the standard deviation of energies given by the latter partition is greater than that given by the former. Along with the significant increase in native conformation energy (from -0.601 to $-0.546kT$), the Z-score of the native conformation increases dramatically (from -2.954 to -2.238). In this case, the Z-score pattern is indicative of threading success (i.e., better threading for the structural partition that gives a higher absolute Z-score value).

General Trends Relating to the Distribution of Threading Scores

To further explain the predictive power of information gain or "energy" in threading, we examine the behavior of the threading spectra as a function of different partitions of sequence and structure spaces. From Figure 5, one can see that the increase of the spectrum width occurs mainly

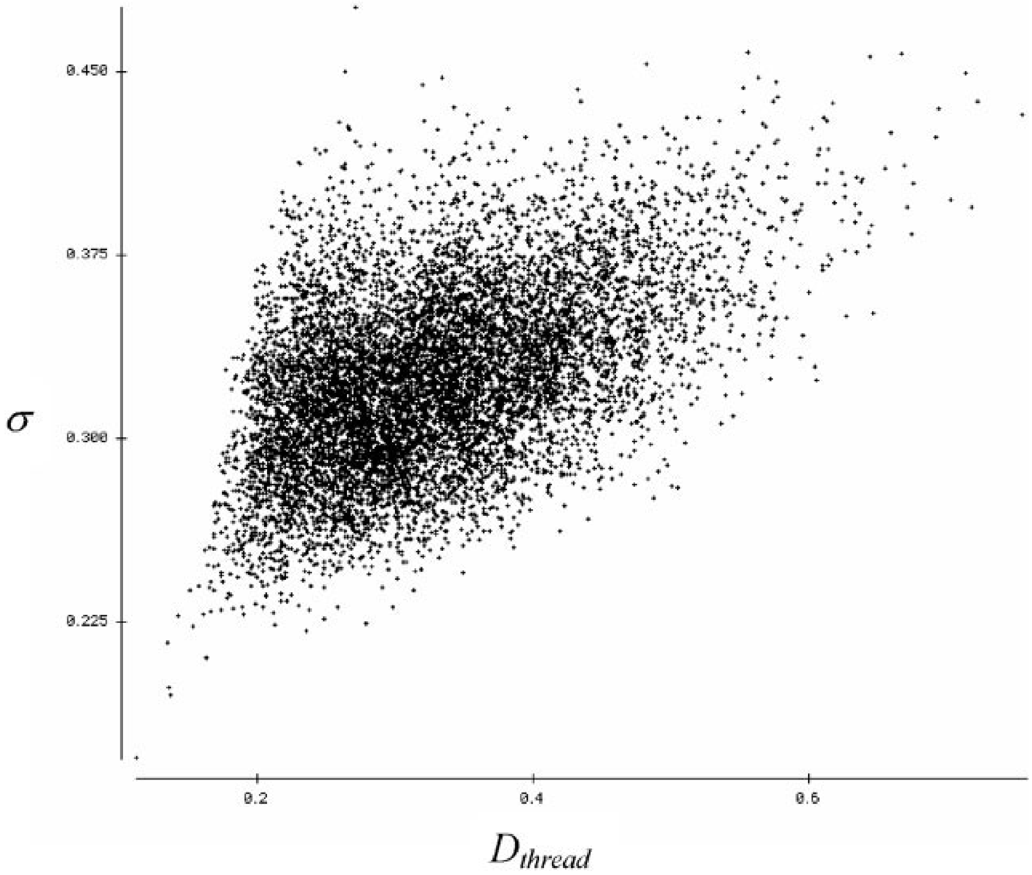

Fig. 7. Behavior of spectral standard deviation and spectral mean. The quantitiy $D_{\text{thread}}(s,C)$, the mean energy of all decoy conformations in the therading pool, is plotted against the standard deviation for the partition $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$. If the sperad in standard deviation is distributed evenly between the two extremes of the spectrum, one should see little correlation between $D_{\text{thread}}(s,C)$ and $\sigma(s,C)$. However, this plot reveals a significant tendency for the standard deviation to increase when the location of the distribution $D_{\text{thread}}(s,C)$ shifts to the right.

at the positive wing of the distribution, while the negative wing remains relatively constant across the three partitions.

This qualitative observation is confirmed by more quantitative methods. As a measure of distribution location, the informatic quantity $D_{\text{thread}}(s,C)$, the mean energy of all decoy conformations in the threading pool, is plotted against the distribution standard deviation for the partition $S_{x\text{-}20\text{-}y}$–$(\phi,\psi)_{20°}$ in Figure 7. If the spread in standard deviation is distributed evenly between the two extremes of the spectrum, one should see little correlation between $D_{\text{thread}}(s,C)$ and $\sigma(s,C)$. But while the scatter plot (Fig. 7) is quite disperse, it reveals a significant tendency for the standard deviation to increase when the location of the distribution $D_{\text{thread}}(s,C)$ shifts to the right. This observation suggests that the widening of the distribution occurs mainly towards the positive extreme of the energy spectrum, while keeping the negative extreme of the spectrum anchored at the same general position. This is typified by the case illustrated in Figure 5.

Grand averages $\langle D_{\text{thread}}(S,C)\rangle$ as a function of different partitions were plotted against $\langle\sigma(s_i,C)\rangle$, and exhibit a similar trend (Fig. 8). Increase in spectrum width is accompanied by an upward shift of the spectrum mean. Further confirmation is shown in Figure 9, in which we plot the average energy values corresponding to a given

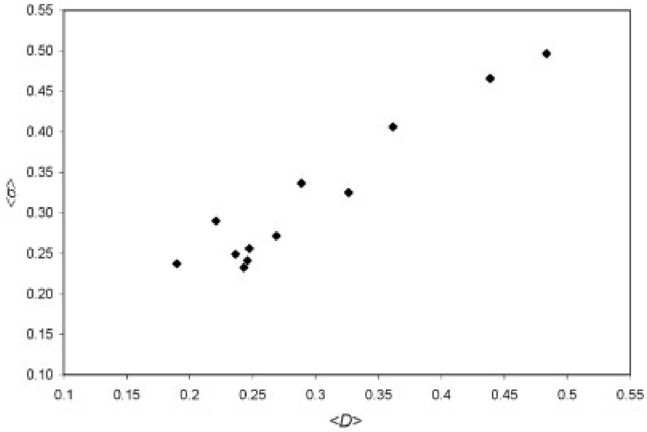

Fig. 8. Behavior of spectral standard deviation $\langle\sigma(s_i,C)\rangle$ against $\langle D_{\text{thread}}(S,C)\rangle$. Each point represents the grand averages computed for sequence-structure partitions studied in this work. Increase in spectrum width is accompanied by an upward shift of the spectrum mean.

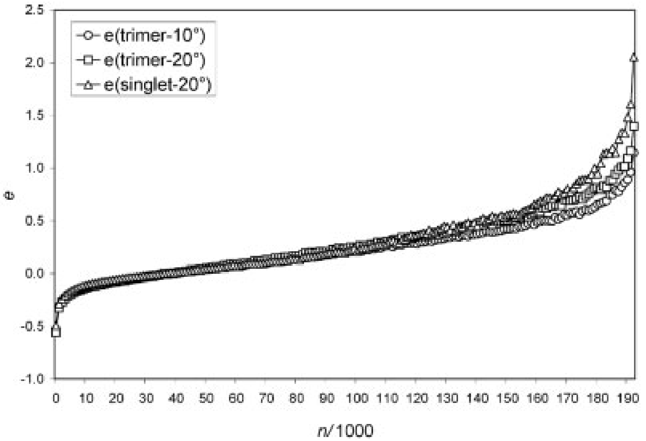

Fig. 9. Average energy values corresponding to a given level of $\langle r_Z(S,C^T)\rangle$ for three different (ϕ,ψ) resolutions. The plot shows little divergence at low $\langle r_Z(S,C^T)\rangle$ values (negative extreme of the energy spectrum). The difference becomes signficiant as one proceeds towards high $\langle r_Z(S,C^T)\rangle$ values.

value of $\langle r_Z(S,C^T)\rangle$, for three different (ϕ,ψ) resolutions (whose distinct energy spectra are pictured in Fig. 5). The plot shows little divergence at low $\langle r_Z(S,C^T)\rangle$ values (negative extreme of the energy spectrum) among the three resolutions, but the difference becomes significant as one proceeds towards the other extreme. In the low-energy region, a comparable level of energy leads to a given amount of discrimination [as measured by $r_Z(s,c^T)$], while in the opposite region, distributions with larger standard deviation are able to tolerate higher energy values to achieve a comparable $r_Z(s,c^T)$, as the spectrum tends to absorb the increased standard deviation in this region.

Manipulation of sequence and structural partitions, which changes information gain (and concomitantly the mean energy of the native conformation), alters the energy spectra of the universe of structures as well, thereby affecting the ability of the potential function to discriminate the native conformation from the pool of incorrect alternative structures. If the low end of the energy spectra remain relatively constant across different partitions, a decrease in native energies should benefit threading discrimination overall, despite its tendency to alter spectral location and width [as measured by $D_{\text{thread}}(s,C)$ and $\sigma(s,C)$, respectively].

CONCLUDING REMARKS

Local Backbone Torsion Potential

The total energy ΔE_{total} is usually derived by a linear combination of the relative contributions ΔE_q of any number of significant interactions, of the general form

$$\Delta E_{\text{total}} = \sum_q \alpha_q \Delta E_q = \sum_q \alpha_q \sum_r \Delta E(c_i|s_j)_{qr} \qquad (15)$$

where the summation over different types of interactions (i.e., local, long-range, electrostatic, solvent interaction, etc.) is modulated by a set of weights α_q, which determine their relative importance. The set of α_qs, which most successfully assigns minimum ΔE_{total} values for the native conformation of each protein in the data set are chosen.

We have focused on the local component (i.e., the backbone torsional potential) of the total energy, and have found that optimizing the parameters used for this component via information maximization increases its effectiveness in identifying native conformation. It is reasonable to assume that the effect of each type of interaction can be enhanced by maximizing its corresponding $I_g(S,C^T)$. If one assures the minimum value for $\sum_r \Delta \mathrm{E}(c_i|s_j)_{qr}$ for each interaction q, ΔE_{total} will take on the minimum value as well. Such a strategy is rooted on the principle of consistency,[31] which asserts that the energetic minima of different types of interactions, taken individually, should, in general, be simultaneously consistent with the native conformation. Further verification of the success of the information maximization strategy as applied to the other components of the energy function (e.g., those used in refs. 9 and 32) will be the subject of future work. We will also investigate the effects of the choice of reference (sequence-free) state, an important factor in building better contact and distance-dependent side-chain–side-chain potentials.[33]

This work outlines a strategy for formulating the best (ϕ,ψ) torsional potential from structural data. We find that including the flanking amino acid residues in the sequence description significantly increases the effectiveness of the potential, compared to using the single-residue description. However, due to the limits of the current data, the description of the flanking residues must be simplified by contracting the amino acid alphabet from the full 20 into a smaller number of clusters, determined automatically by the information maximization methodology.

The Informatic Basis of Statistical Potentials and Empirical Threading Score Functions

We have shown here that statistical potentials and threading score functions, derived from finite data sets, are equivalent to informatic functions, and are therefore dependent on the way data are classified and simplified. Although complexities in the sequence and structural domains of protein chains must be reduced to make use of limited experimental data, oversimplification obscures relationships between sequence and structure that are critical in structure prediction. Optimization of information extraction[2,3] by manipulation of sequence and structural descriptors (alphabet contraction, sequence window size, backbone descriptor choice, and structural resolution), involves compromise between the need to describe data in detail and the limited size of the data set itself.

The choice of sequence and structural parameters necessarily affects estimates of conditional probabilities $p(c|s)$, which determine the amount of information extracted from the data set, as measured by the information gain. The mathematical relationship between information gain and mean energy, established in this work, demonstrates that manipulation of descriptive parameters also alters the energy values assigned to native conformation, and consequently, the performance of statistical potential functions in threading. Sequence and structural partitions that maximize information gain minimize the mean energy of the ensemble of native conformations.

The effect on threading success is a more complex matter, as it involves the comparison of the energy or score of the native conformation with the ensemble of incorrect sequence-structure matches, or the entire energy spectrum of the universe of conformations. In this work, we have established an informatic basis for the placement of the native score within an energy spectrum. Although the derived quantities, $\langle \Delta E(C^T|S) \rangle$, $\langle D_{\text{thread}}(S,C) \rangle$, $\langle Z(S,C^T) \rangle$, and $\langle \sigma(s_i,C) \rangle$ are all grand averages derived from many threading operations (in fact, all possible threading operations), they should give an indication as to the likely behavior of an individual threading. We find that among these quantities, the mean energy $\langle \Delta E(C^T|S) \rangle$ is the best predictor of threading success, as measured by two quantities: (1) the likelihood of assigning the lowest energy to the native conformation and (2) the mean ranking of the native conformation energy among the pool of alternative conformations, or $\langle r_Z(S,C^T) \rangle$. The mean Z-score (either the standard mean or the harmonic mean), a standard measure of discrimination,[18,34] does not always predict the behavior of a statistical potential in threading. This is due to the fact that Z-score is dependent on the standard deviation of the spectrum of energy scores of the pool of decoy structures. As shown in the example illustrated in Figure 5, a substantial increase in standard deviation may result in a lower Z-score, but the position of the native score relative to those given by decoy structures may actually be closer to the left extreme of the spectrum. Therefore, building potentials by optimizing Z-score may not always improve performance in fold recognition.

The predictive value of $\langle \Delta E(C^T|S) \rangle$ in threading suggests that $I_g(S,C^T)$ must also be a good indicator of threading success. This implies that any parameter that increases information gain should also improve the threading performance of a statistical potential derived from the same probability distributions. Based on our results, the choices in descriptors which maximize information gain must also produce the best potential functions. Strategies to optimize these parameters with respect to information extraction are therefore relevant to the effectiveness of statistical potentials in the major effort to computationally predict structure from sequence.

REFERENCES

1. Rackovsky S. On the nature of the protein folding code. Proc Natl Acad Sci USA 1993;90:644–648.
2. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins Struct Funct Genet 2000;38:149–164.
3. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins Struct Funct Genet 2002; 48:463–486.
4. Lambert MH, Scheraga HA. Pattern recognition in the prediction of protein structure. II. Chain conformation from a probability-directed search procedure. J Comput Chem 1989;10:798–816.
5. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
6. Lee B, Kurochkina N, Kang HS. Protein folding by a biased Monte Carlo procedure in the dihedral angle space. FASEB J 1996;10: 119–125.
7. Gibbs N, Clarke AR, Sessions RB. Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. Proteins Struct Funct Genet 2001;43:186–202.
8. Sippl MJ. Calculation of Conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
9. Sun S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci 1993;2:762–785.
10. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
11. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
12. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins Struct Funct Genet 1999;36:54–67.
13. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. J Mol Biol 1990;216:167–180.
14. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Mol Des 1993;7:473–501.
15. 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.
16. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
17. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins Struct Funct Genet 2001;44:223–232.
18. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
19. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628–11633.
20. Bienkowska JR, Rogers RG, Smith TF. Performance of threading scoring functions designed using new optimization method. J Comput Biol 1999;6:299–311.
21. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins Struct Funct Genet 2000;41:40–46.
22. Meller J, Elber R. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins Struct Funct Genet 2001;45:241–261.
23. Chhajer M, Crippen GM. A protein folding potential that places the native states of a large number of proteins near a local minimum. BMC Struct Biol 2002;2:4–14.
24. Kuznetsov IB, Rackovsky S. Discriminative ability with respect to amino acid types: assessing the performance of knowledge-based potentials without threading. Proteins Struct Funct Genet 2002; 49:266–284.
25. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins Struct Funct Genet 1999;36:357–369.
26. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
27. 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.
28. Kullback S. Information theory and statistics. New York: Wiley; 1959.
29. Vendruscolo M, Mirny LA, Shakhnovich EI, Domany E. Comparison of two optimization methods to derive energy parameters for protein folding: perceptron and Z-score. Proteins Struct Funct Genet 2000;41:192–201.
30. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
31. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:193–210.
32. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
33. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
34. Sippl MJ, Jaritz M, Hendlich M, Ortner M, Lackner P. Applications of knowledge based mean fields in the determination of protein structures. In: Doniach, S., editor. Statistical mechanics, protein structure, and protein substrate interactions. New York: Plenum Press; 1994.