PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Information and discrimination in pairwise contact potentials

Armando D. Solis and S. Rackovsky*

Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029

ABSTRACT

We examine the information-theoretic characteristics of statistical potentials that describe pairwise long-range contacts between amino acid residues in proteins. In our work, we seek to map out an efficient information-based strategy to detect and optimally utilize the structural information latent in empirical data, to make contact potentials, and other statistically derived folding potentials, more effective tools in protein structure prediction. Foremost, we establish fundamental connections between basic information-theoretic quantities (including the ubiquitous Z-score) and contact "energies" or scores used routinely in protein structure prediction, and demonstrate that the informatic quantity that mediates fold discrimination is the total divergence. We find that pairwise contacts between residues bear a moderate amount of fold information, and if optimized, can assist in the discrimination of native conformations from large ensembles of native-like decoys. Using an extensive battery of threading tests, we demonstrate that parameters that affect the information content of contact potentials (e.g., choice of atoms to define residue location and the cut-off distance between pairs) have a significant influence in their performance in fold recognition. We conclude that potentials that have been optimized for mutual information and that have high number of score events per sequence–structure alignment are superior in identifying the correct fold. We derive the quantity "information product" that embodies these two critical factors. We demonstrate that the information product, which does not require explicit threading to compute, is as effective as the Z-score, which requires expensive decoy threading to evaluate. This new objective function may be able to speed up the multidimensional parameter search for better statistical potentials. Lastly, by demonstrating the functional equivalence of quasi-chemically approximated "energies" to fundamental informatic quantities, we make statistical potentials less dependent on theoretically tenuous biophysical formalisms and more amenable to direct bioinformatic optimization.

Proteins 2008; 71:1071–1087.

Key words: pairwise contact potentials; empirical potentials; statistical potentials; information theory; protein fold recognition; threading; divergence.

INTRODUCTION

Numerous contacts occur in folded proteins between amino acid residues that are far apart in sequence. The patterns of these so-called long-range interactions have been utilized in protein structure prediction as a criterion to judge the correctness of a given sequence–structure alignment. The scoring scheme is embodied in energy-like empirical potentials,[1–4] derived easily from the database of solved high-resolution X-ray structures.[5] The success of these contact potentials, along with their relative ease of application, has made them an important part of statistical potentials for fold recognition and ab initio prediction.[6–9]

The effort to design better folding potentials seeks to unlock as much of the structural information encoded in sequence as possible. It is an axiom of protein folding that all the information needed to determine the structure of a protein chain is contained in its amino acid sequence. Therefore, the remarkable performance of contact potentials in discriminating between native and non-native conformations implies that pairwise contacts between residues contain a significant amount of this information. Understanding the nature of long-range pairwise contact information, both in the way it is encoded in sequence and its ability to discriminate among native-like conformations, is of primary interest for optimizing performance. In this work, we use information-theoretic ideas, developed in previous work,[10] to carry out two critical analyses: first, to quantify the information residing in pairwise contacts derived from known structure, and then to understand the behavior of contact potentials in fold recognition as a function of this information.

Two previous studies[11,12] have concluded that the mutual information contained in pairwise contacts, as estimated from observed propensities, are modest at best. The question arises how such a minute amount of information can lead to the proven success of contact potentials in fold recognition. Our work attempts to reconcile these two seemingly divergent observations. We first note that the two studies did not explic-

Grant sponsor: National Library of Medicine of the National Institutes of Health; Grant number: LM006789.
*Correspondence to: S. Rackovsky, Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, Box 1603, One Gustave L. Levy Place, New York, NY 10029. E-mail: shelly@camelot.mssm.edu
Received 27 December 2006; Revised 16 June 2007; Accepted 21 June 2007
Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21733

itly optimize potential parameters before measuring mutual information. Here, we demonstrate that the definition of the interaction and the parameterization of the potential function have a significant effect on the amount of mutual information, in line with our previous work on sequence and structure representation.[13,14] In fact, we have found significant variance in the amount of information as a function of these factors, and a higher mutual information for particular definitions and parameters.

Whatever the level of information measured, a more fundamental question arises: is mutual information the quantity responsible for the discrimination of native conformation among a large ensemble of incorrect or "decoy" conformations? In this work, we address this issue by exploring the quantities that define the "informational landscape" of folding. We find that discrimination is described more accurately by total divergence, an informatic quantity that includes, as one of two components, the mutual information. It is through total divergence that mutual information affects discrimination.

Simultaneously, we advance an informatic description of statistical potentials, to replace the cumbersome energetic (or pseudo-energetic) description, which relies on theoretically tenuous biophysical justifications.[15–18] Among empirical methods, the probabilistic dissection of score functions[10,19–24] has major advantages.[18] While the classical or pseudo-"energies" and probability and information "scores" are quantitatively similar, viewing the action of potentials through the prism of probability and information theories is beneficial because of their flexibility to incorporate many factors and interactions, without the need to find explanations consistent with complex biophysical reality. Furthermore, tools to maximize information become available for the optimization of the score functions.[10] This does not, however, mean that unphysical parameters or potential configurations will be found to be optimum. The automatic use of information maximization has continued to yield biophysically consistent descriptors, such as the optimal clustering of similarly coding amino acids.[14]

This article accomplishes four analytical procedures: (1) we derive the informatic quantities that describe the "energy" of the sequence-conformation alignment (i.e., gapless threading), as well as the discrimination between native and decoy conformations; (2) we explore the typical behavior of a contact potential in a threading exercise; (3) we survey the effect of adjustable parameters and two specific definitions of contact (i.e., distance of C_β atom pair and closest approach of heavy atom pair) on the informatic quantities and fold recognition performance; (4) the availability of a large body of threading data allows us to make a direct link between the informatic quantities and performance.

We find that the influence of parameters that characterize statistical potentials on threading performance fall under two overlapping categories: those that affect the amount of information conveyed by the potential and those that affect the number of score events per sequence-structure alignment. The former can be measured by direct information-theoretic means.[10] In the case of contact potentials, the latter refers to the number of contacts observed when a given sequence is mounted onto a trial conformation. We show that this quantity measures the stability of mutual information (an *averaged* quantity) to indicate success in the potential's application to individual sequences. As will be explored in detail here, we find that higher mutual information and a large number of score events are crucial to improve the performance of potentials. We derive a quantity to embody these conclusions, which we call the "information product." It performs as well as the ubiquitous *Z*-score (a related information-theoretic quantity) in indicating how well a potential will perform in fold recognition, but with a crucial difference: while an extensive battery of threading sequences through decoys is required by the mean *Z*-score,[25] the information product can be measured simply by a survey of sequences in their native conformations, without need for trial threadings of large decoy ensembles.

The main thrust of this and previous work has been to provide a direct link between information and performance of statistical potentials. Through the detailed study of contact potentials, this work further advances the idea that maximizing information encoded in potentials is a viable way to improve their over-all ability to identify the correct fold amidst large ensembles of possible conformations.

THEORY

In an effort to properly measure the relationship between information and performance, we investigate the effect of applying a given potential configuration across a range of protein folds and sequences. The quantities of interest are therefore *averages* of measurements taken from a series of gapless threading exercises using a large and diverse set of folds. The best measurements that can be derived from a large set of nonredundant protein sequences are those that arise from exhaustive threading of *all* sequence segments (of set lengths) through *all* possible native conformations that exist in the data set. Not coincidentally, informatic quantities that can characterize a given potential arise naturally from such an all-against-all threading operation.

We first summarize information-theoretic concepts, similar to the analysis in previous work on local backbone potentials.[10] We then establish an informatic basis for ordinary contact energy scores. Once "energies" are shown to be classical informatic quantities, we can analyze the scores from a general information-theoretic viewpoint.

Mutual information and divergence

The amount of information contained in one random variable about another can be measured by mutual information:

$$I(X;Y) = \sum_{(x,y)} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where $p(x,y)$ is the joint probability distribution of X and Y, two variables of interest, and $p(x)p(y)$ is the product of their individual distributions. If X and Y are uncorrelated, $p(x,y) = p(x)p(y)$, and therefore $I(X;Y) = 0$. Otherwise, $I(X;Y) > 0$, conforming to our notions of how information behaves. We refer the reader to Ref. 26 for a detailed discussion of these properties, and for further study of information theory.

Mutual information is a special instance of divergence, a fundamental informatic quantity, which measures the distance between two probability distributions. The divergence between two probability distributions $p(z)$ and $q(z)$ is

$$D[p||q] = \sum_z p(z) \log \frac{p(z)}{q(z)} \quad (2)$$

It can be seen that in the special case of $Z = (X,Y)$, with $q(z) = p(x)p(y)$, the divergence equation reduces to mutual information. This quantity is also known as directed divergence, relative entropy, and Kullback–Leibler distance. For our purposes, it is also useful to think of divergence as a measure of inaccuracy in using a distribution $q(z)$ to describe an event rather than its correct distribution $p(z)$. This will be especially useful in thinking about the reference state of a potential function, which is the expected distribution of conformations before the folding force field is introduced.

Because $D[p\|q] \neq D[q\|p]$, divergence is not strictly a distance measure. However, the divergence measure can be made symmetric by the following operation:

$$J[p,q] = D[p||q] + D[q||p] \quad (3)$$

The quantity J, called the total divergence, is a measure of the total discrimination of the two probability distributions. A link has been established between J and common scoring functions for backbone conformations used for fold recognition.[10] Here, we demonstrate that a similar relationship exists between informatic functions and scores arising from the pairwise contact potentials.

The contact potential

The contact energy $E^T_{C|S}$ of a sequence S in a conformation C is usually defined as the sum of individual scores of all amino acid pairs (x,z) judged to be in physical contact:

$$E^T_{C|S} = \sum\nolimits^{S}_{(i,j)} \mathbf{i}_{i,j|C}(d_0)\mu^{ij}_{xz} \quad (4)$$

where $\mathbf{i}_{i,j|C}(d_0)$ is the contact map index specific to C, which takes a value of 1 if the distance between residues in positions i and j is within a specified cut-off distance (i.e., $d^{ij} \leq d_0$, where d_0 is the maximum contact distance) and 0 otherwise. The distance d_0 is a critical variable. Since the contact map $\mathbf{i}_{i,j|C}(d_0)$ depends on the definition of a contact, the contact energy $E_{C|S}$ also varies with d_0. The notation μ^{ij}_{xz} represents the contact score of the amino acid pair (x,z) located in positions i and j along the chain of sequence S. One way to derive the contact score is to use the pseudo-energy framework. The score takes the form

$$\mu_{xz} = -\log \frac{p(a_x a_z)^{obs}}{p(a_x a_z)^{exp}} \quad (5a)$$

where $p(a_x a_z)^{obs}$ is the observed probability of contact between amino acids a_x and a_z, and $p(a_x a_z)^{exp}$ is the expected contact probability, based mainly on frequencies of the amino acids in sequence space. Called the reference distribution, this quantity is central to the effectiveness of the energy function.[27–30] In this study, we restrict ourselves to the simplest quasi-chemical form:

$$p(a_x a_z)^{exp} = p(a_x)p(a_z) \quad (5b)$$

where $p(a_j)$ is the frequency of amino acid j in the (nonredundant) protein structural database. Though tracing the informatic roots of contact potentials is most straightforward with this choice, the information equations derived in this work should be generalizable to other kinds of reference states. (An informatic analysis of reference states will be the subject of a future publication.) The quantity $p(a_x a_z)^{obs}$ is the frequency of contact between the two amino acids observed in a representative ensemble of native conformations. To indicate the source of the statistic, we rewrite this quantity as $p(a_x a_z)_N$, that is, the contact frequency of the pair given natural folded conditions.

Link between pseudo-energy contact scores and mutual information

The energy per contact for a sequence S in its native conformation N is

$$E_{N|S} = \frac{1}{n_t} \sum\nolimits^{S}_{(i,j)} \mathbf{i}_{i,j|N}(d_0)\mu^{ij}_{xz} \quad (6)$$

where n_t is the number of contacts (or the number of 1s in the contact map matrix, or more generally, the number of score events), with the summation covering all

(i,j) position pairs of sequence S. The average energy of a typical sequence in its native conformation N is

$$\begin{aligned}\langle E_{N|S}\rangle &= \frac{1}{n_S}\sum_S E_{N|S} \\ &= \frac{1}{n_S}\sum_S \frac{1}{n_{t(N|S)}}\sum\nolimits^S_{(i,j)} \mathbf{i}_{i,j|N}(d_0)\mu^{ij}_{xz} \\ &= -\frac{1}{n_S}\sum_S \frac{1}{n_{t(N|S)}}\sum\nolimits^S_{(i,j)} \mathbf{i}_{i,j|N}(d_0)\log\frac{p(a^i_x a^j_z)_N}{p(a^i_x)p(a^j_z)}\end{aligned} \tag{7a}$$

where $n_{t(N|S)}$ is the total number of contacts of sequence S in its native conformation. Eq. (7a) is equivalent to finding the expected value of the log-odds score μ_{xz}. Another way to express this expectation is to collate all instances in the data set of every *unique* amino acid pair (x,z) in contact, and then recast the equation as a summation over all unique amino acid pairs (x,z), instead of position pairs (i,j). Normalizing the number of contacting (x,z) pairs in the data set by $1/n_S$ and $1/n_{t(N|S)}$ (i.e., the total number of contacts irrespective of sequence) converts the count into the frequency $p(a_x a_z)_N$. The equation can be rewritten thus:

$$\langle E_{N|S}\rangle = -\sum_{(x,z)} p(a_x a_z)_N \log\frac{p(a_x a_z)_N}{p(a_x)p(a_z)} \tag{7b}$$

where the summation is over all unique amino acid pairs. Therefore, using Eq. (1),

$$\langle E_{N|S}\rangle = -I(a_x, a_z) \tag{8a}$$

According to this relation, the information contained in amino acid contact patterns can be measured directly from the average per contact "energy" of native conformations. We note that the average contact energy can also be cast as a divergence, via Eq. (2):

$$\langle E_{N|S}\rangle = -D[(a_x a_z)||(a_x)(a_z)] \tag{8b}$$

Through this equivalence between log-odds potential functions and mutual information, we also find that increasing the efficiency of information extraction changes the average per contact score of the sequence in its native conformation. Moreover, the search for the "energy minima" can be recast in bioinformatic terms as a search for optimal use of sequence-dependent structural information.

Link between gapless threading score and total divergence

Searching for the global energy minimum involves not only knowing the score of the correct sequence–structure alignment but also evaluating the energies of the entire ensemble of possible conformations. Given a sequence S, the per contact score of the native conformation can be compared to the average score generated, in a gapless threading exercise, by the set $\{C\}$ of non-native conformations. A possible objective function to measure discrimination is the gap between the native score and the mean score given by the ensemble of decoy conformations:

$$\begin{aligned}E_{N|S} - \langle E_{C|S}\rangle = &\left\{\frac{1}{n_{t(N|S)}}\sum\nolimits^S_{(i,j)} \mathbf{i}_{i,j|N}(d_0)\mu^{ij}_{xz}\right\} \\ &- \left\{\frac{1}{n_C}\sum_{\{C\}}\frac{1}{n_{t(C|S)}}\sum\nolimits^S_{(i,j)} \mathbf{i}_{i,j|C}(d_0)\mu^{ij}_{xz}\right\}\end{aligned} \tag{9}$$

(We note that the Z-score, a more common measure of discrimination, shares some basic similarities to this function. We will discuss this issue in a later section.) This discrimination gap can be measured for all sequences in the data set, resulting in an expected gap value:

$$\langle E_{N|S} - \langle E_{C|S}\rangle\rangle = \langle E_{N|S}\rangle - \langle\langle E_{C|S}\rangle\rangle \tag{10}$$

The informatic meaning of the first term has been examined in a previous section [Eq. (8)]. We examine the second term:

$$\langle\langle E_{C|S}\rangle\rangle = \frac{1}{n_S}\sum_S \frac{1}{n_C}\sum_{\{C\}}\frac{1}{n_{t(C|S)}}\sum\nolimits^S_{(i,j)} \mathbf{i}_{i,j|C}(d_0)\mu^{ij}_{xz} \tag{11a}$$

This term expresses the expected value of μ_{xz}, averaged over all possible sequence *and* conformation combinations. The three summations in Eq. (11a), in conjunction with the contact matrix index, describe an accounting of (x,z) pair contacts in the data set of *all* sequences mounted onto *all* possible native-like conformations in structural space. With the three normalizing factors (n_S, n_C, and $n_{t(C|S)}$), the operation is equivalent to finding the expected frequency of contact between every pair of amino acids in folded chain configurations. With the assumption that amino acid residues are distributed randomly across protein sequences, the expected frequency of contact in decoy threading can be shown to be dependent only on their relative populations.[31] This leads to a simpler expression, involving a summation over all unique (x,z) pairs:

$$\begin{aligned}\langle\langle E_{C|S}\rangle\rangle &= \sum_{(x,z)} p(a_x)p(a_z)\mu_{xz} \\ &= -\sum_{(x,z)} p(a_x)p(a_z)\log\frac{p(a_x a_z)_N}{p(a_x)p(a_z)}\end{aligned} \tag{11b}$$

which can be written as a divergence:

$$\langle\langle E_{C|S}\rangle\rangle = D[(a_x)(a_z)||(a_x a_z)] \tag{11c}$$

Therefore, the expected value of discrimination is the total divergence:

$$\langle E_{N|S} - \langle E_{C|S}\rangle\rangle = -D[(a_x a_z)||(a_x)(a_z)] - D[(a_x)(a_z)||(a_x a_z)] = -J[(a_x a_z),(a_x)(a_z)] \quad (12a)$$

While the reference stated in this derivation is the simple product of frequencies of residue pairs, these equations can be extended to other valid types of reference states, so long as they satisfy the reduction from Eq. (11a) to Eq. (11b). For now, we state that the generalized form of the threading objective function, in the framework of information theory, is

$$\langle E_{N|S} - \langle E_{C|S}\rangle\rangle = -D[(a_x a_z)_N||(a_x a_z)_R] - D[(a_x a_z)_R||(a_x a_z)_N] = -J[(a_x a_z)_N,(a_x a_z)_R] \quad (12b)$$

where N refers to the native folded state, and R refers to the reference state. Discrimination of the native conformation is quantified by the distance between the native and reference distributions of pair contacts. This is a generalized relationship between the score given by a potential (of any kind) and its ability to discriminate, which we derived first for the specific case of the local sequence dependence of backbone conformation.[10] We have therefore demonstrated the equivalence between "energetic" quantities used in contact-threading studies and well-defined informatic functions. In the present work, we apply these results to the information-theoretic analysis of the contact potentials.

METHODOLOGY

The informatic equations derived above provide a framework within which to analyze the information content in contact potentials, with respect to their performance in fold recognition. The initial step is to measure the information contained in contact potentials. We then analyze the relationship between the definition of a contact and the potential's information content. Next, we establish a rigorous method to measure the effectiveness of contact potentials in fold recognition and examine the informatic factors that affect their performance. In this section, we outline the mechanics of the threading procedure, methods for computing the informatic functions, and the measures of threading performance.

Gapless threading score

The link we have established between $E_{C|S}$ scores and informatic functions enables us to abandon the "energy" formalism in favor of information theory.

In place of $E_{N|S}$, the contact energy of sequence s, of length L, in its correct conformation is defined as the specific information:

$$I_L^*(s) = -\frac{1}{n_{t(N|S)}}\sum\nolimits_{(i,j)}^{s} \mathbf{i}_{i,j|N}(d_0)\mu_{xz}^{ij} \quad (13a)$$

The mean energy of all decoy conformations is the specific divergence:

$$D_L^*(s) = \frac{1}{n_C}\sum\nolimits_{\{C\}}\frac{1}{n_{t(C|S)}}\sum\nolimits_{i,j}^{S} \mathbf{i}_{i,j|C}(d_0)\mu_{xz}^{ij} \quad (13b)$$

The difference between these two quantities is the discrimination function, the specific total divergence:

$$J_L^*(s) = I_L^*(s) - D_L^*(s) \quad (13c)$$

These three functions are the specific-case analogs of the classic informatic functions I, D, and J, which are their expectations, or

$$E[I_L^*(s)] = I(a_x, a_z) \quad (14a)$$

$$E[D_L^*(s)] = -D[(a_x)(a_z)||(a_x a_z)] \quad (14b)$$

$$E[J_L^*(s)] = J[(a_x a_z),(a_x)(a_z)] \quad (14c)$$

The negative sign in Eq. (14b) is needed to conform to the definition of divergence [Eq. (2)]. In particular, it justifies the form taken by the total divergence in Eq. (13c). With the negative sign, the total divergence turns into a sum of an information and a directed divergence, which is the expected functional form of J [as defined in Eq. (3)]. We also note that since these functions measure information (as opposed to "energy"), desirable scores for native conformations are positive. Therefore, the objective of discrimination is to assign a *higher* score to the native conformation than any other conformation.

Gapless threading procedure

To systematically evaluate factors of interest, we carry out a series of threading exercises aimed at examining the behavior of the scoring function across different sequences and conformations. We work with a nonredundant set of sequences and their associated conformations, derived from the PDBselect database by Hobohm and Sander[32] (http://bioinfo.tg.fh-giessen.de/pdbselect/). We selected the high-resolution structures of 1036 protein chains whose sequences share no more than 25% sequence similarity. The resulting data set has a total of 210,995 amino acid residues.

Our object is to evaluate the behavior of informatic and threading quantities given a particular definition of energy or score. We examine specific factors that can affect the score, including the length of the amino acid sequence (L), the contact cut-off distance (d_0), and which atoms are used to determine distance between residues. Distinct score functions are defined by different combinations of values of the latter two factors.

A primary concern is the atoms chosen to determine whether a residue pair is in contact. We investigate two cases here. The distance between two residues in the heavy-atom (HA) approximation is given by the closest approach of heavy atoms (i.e., nonhydrogen atoms) in the residues (side chain and main chain atoms). As an alternative, one can choose the position of a specific atom to indicate the location of each residue. In this work, in addition to the HA scheme, we consider the CB scheme, which uses the distance of C_β atoms to define the distance between a residue pair. To eliminate undue effect of chain connectivity, pairs within five (for HA) or eight (for CB) residue positions on the chain are not included in the contact statistics, or in scoring a sequence–structure alignment.

The cut-off distance, d_0, is another critical parameter in the definition of interresidue contact. For both the HA and CB schemes, we build different potentials by varying d_0 within the range of 4.0–6.5 Å in HA and 8.5–13.5 Å in CB.

To examine the effect of sequence length on the performance of contact potentials, we examine five different sequence lengths: 50, 100, 150, 200, and 250 residues. Each continuous protein chain in the data set is broken down into overlapping chains of the specified length, while chains that are shorter than the length scale being considered are dropped from the analysis. For instance, a protein chain of 175 residues (without any chain break) will yield 126 overlapping 50-mers, 76 100-mers, and 26 150-mers, but will be dropped from analysis of lengths 200 and 250. These L-mers can then be used as amino acid sequences to be threaded, and their structures can be included in the decoy ensemble on which to thread other sequences.

A threading exercise involves the comparison of the native score of a given sequence with the average score given by an ensemble of test conformations derived from the data set. To measure trends for a particular energy function in the best way, and to conform to the derivation of the informatic quantities in the previous section, *all* possible L-mer sequences in the data set are threaded through *all* L-mer conformations in the data set. This all-against-all operation yields the most comprehensive battery of threading possible given a structural data set. The total number of L-mers in the data set for $L = 50$, 100, 150, 200, and 250 are, respectively, 137,328, 89,625, 58,034, 37,727, and 24,083. This means that for $L = 50$, 137,328 50-mer sequences are tested, with each sequence threaded onto its native conformation and 137,327 decoys. The scores for each sequence–structure alignment are computed using the given potential function. These measurements facilitate the computation of grand average scores, corresponding to the informatic functions of interest: mutual information I, directed divergence D, and total divergence J.

We note that partitioning the sequence into sequence segments may seem unphysical, since the information from the *entire* sequence is necessary fold into native conformations, composed of a well-defined hydrophobic core enclosed by a polar surface. However, in this study, we are concerned only with the contacts between residues, and not their relative placement with respect to the aqueous environment. If one imagines a given sequence trying out different conformations in the search for its native structure, isolating a sequence segment, and following its particular search, simulates our methodology. While interactions with parts of the sequence not within the given segment are not counted in the final score, it is certainly desirable that whatever interactions occur within the sequence segment contribute favorably to the native score. Moreover, potentials also work to rule out "high-energy" conformations, which we believe is more rigorously tested by subjecting sequences to the largest ensemble of alternative conformations as possible. Implicit in this analysis is the principle of minimum frustration,[33] which empowers the segmentation of the total energy into individually optimizeable quantities. In our view, the grand averages derived from an extensive all-against-all threading of sequence segments adequately gauges the fitness of a contact potential in the action of selecting correct over incorrect contacts in the full spectrum of all possible pair orientations. We hold that a contact potential that does well in folding sequence segments ought to do well in folding the entire sequence. As an aside, we find it remarkable that sequence segments as short as 50 residues are remarkably successful in identifying its native contact map to within the top 3rd-percentile (as will be discussed in the Results section below).

Evaluating the effectiveness of contact potentials

For each L-mer sequence, a native score $I_L^*(s)$ is computed using the correct conformation, along with an ensemble of scores given by alignment with decoy structures. The rank of the native score, $r(s)$, of sequence s is determined by comparison with the ensemble. This measure allows the most direct evaluation of the discriminatory power of the given potential function. An $r(s)$ value of 1 signifies that the native score is the best over-all. To locate the native rank in the Gaussian distribution of scores, we normalize the rank with the total

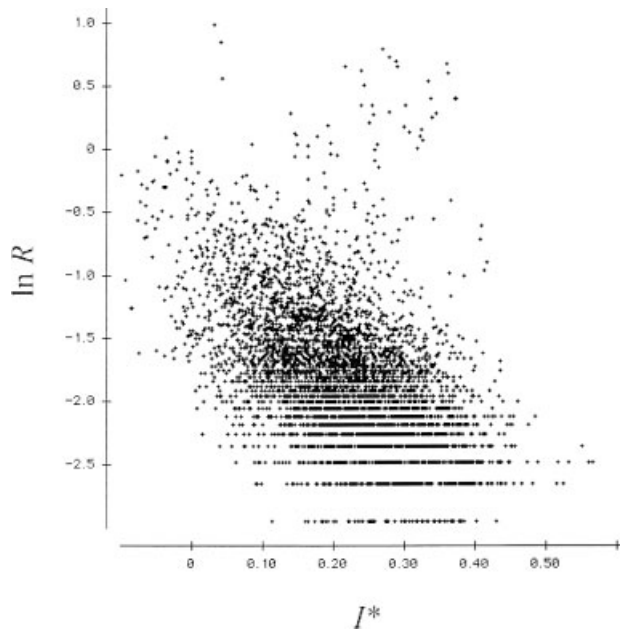

Figure 1

The relationship between individual native contact scores (I) and their percentile rank (R) in an ensemble of decoy conformations. In this illustration, five-thousand randomly picked 100-mer sequences were threaded onto 89,625 decoy conformations, and the contact scores of their native conformations, as assigned by the HA contact potential with $d_0 = 4.5$ Å, were ranked with the scores from the decoy ensemble to give the percentile rank. The correlation is only moderate (Pearson product-moment correlation = −0.496).*

number of conformations tested, $n_{L\text{-mer}}$, to yield the percentile rank:

$$R(s) = \frac{r(s)}{n_{L\text{-mer}}} \times 100 \qquad (15)$$

where the factor 100 is included to return a percentile value. The mean percentile rank can be computed from comprehensive threading:

$$\langle R \rangle = \frac{1}{n_{L\text{-mer}}} \sum_{\substack{\text{all} \\ \text{seqs}}} R(s) \qquad (16)$$

Another way to measure the effectiveness of contact potentials is to actually determine the proportion of sequences tested whose native conformation is assigned the best:

$$R_0 = \frac{n(r=1)}{n_{L\text{-mer}}} \times 100 \qquad (17)$$

where $n(r = 1)$ is the number of times the rank is 1 out of $n_{L\text{-mer}}$ total number of sequences. The factor 100 is included, again, to convert the frequency into a percent value. An R_0 value of 100% means that all the possible L-mers have been successfully assigned the best score by the contact potential. We note that this measure can only be used to compare performance of contact potentials in threading when the total number of decoy structures is the same. It is clear that the likelihood of achieving rank 1 is greater if the number of decoys in the structure ensemble is smaller. Because the range of decoy ensemble size is significant (24,083–137,328) as a function of sequence length, the measure R_0 is useful only within each length. For a more comprehensive evaluation of performance of *all* potentials, we use $\langle R \rangle$, which is not significantly biased by sample size issues.

RESULTS AND DISCUSSION

Comprehensive all-against-all threading was carried out for five sequence lengths (L = 50, 100, 150, 200, and 250), and for a range of cut-off contact distances d_0, for both heavy-atom (HA) or C_β distance (CB) definitions of contact. We begin this section with a discussion of the mechanics of threading using a typical contact potential. As an illustration of threading behavior, we look at the results for a particular potential. We then move on to the discussion of the over-all behavior of each contact potential in comprehensive threading, and make the connec-

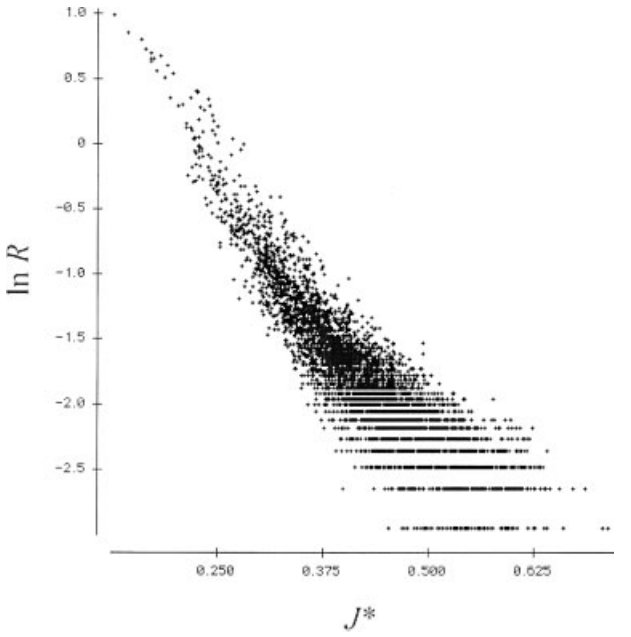

Figure 2

The relationship of the score gap (J) between the native score (I*) and the mean score of decoy conformations (D*) with their percentile rank (R) in an ensemble of decoy conformations. The same 5000 randomly picked 100-mer sequences used for illustration were subject to the threading process, as described in Figure 1. The correlation is strong (Pearson product-moment correlation = −0.885).*

tion between the informatic functions and threading performance, to assist in identifying factors that significantly affect the action of statistical potentials in native fold discrimination.

Information-related details of threading results

We begin by exploring the results from one particular contact potential. We choose the HA potential, with $d_0 = 4.5$ Å and a sequence length of 100 residues, as a representative example. There are 89,625 100-mer sequences in the data set, which means that in comprehensive all-against-all threading, 89,625 sequences are threaded onto the same number of conformations. For each sequence, one of the conformations in the ensemble is native. That alignment gives the native score. The rest of the conformations in the ensemble give "decoy" scores, which are pooled to construct the spectrum of scores used to evaluate the effectiveness of the native score.

Figure 1 indicates that the specific mutual information, $I_L^*(s)$, is only moderately correlated with threading success *for individual cases*. $I_L^*(s)$ of the native conformation can take on a wide range of values, yet can still be the best score as compared to the ensemble of scores given by the decoys. Figure 2, which shows a better correlation between the specific total divergence $J_L^*(s)$ and native score rank, illustrates this idea. It is not the absolute value of the specific information, but its value relative to the mean score of incorrect structures, that is more indicative of threading rank. We note that the behavior of the specific information $I_L^*(s)$ stands in contrast to that of the mutual information I, its expected value, as will be seen in the section below. We also observe the logarithmic relationship between $J_L^*(s)$ and R. Specifically, even a slight decrease (or increase) in J can produce a significant increase (decrease) in the rank of the native conformation. This last property will prove crucial in identifying performance-determining factors, which we explore in a later section.

Information in contact potentials

We now turn our attention to the over-all behavior of contact potentials. The average threading behavior of the potentials studied here, over a range of cut-off distances d_0 in both the closest heavy-atom (HA) and C_β distance (CB) definitions, are summarized in Tables I and II.

Our results indicate that contact potentials can encode variable amounts of information, depending on parameterization, and can perform better than what previous measurements[11,12] suggest. Mutual information ranges

Table I
Informatic Properties and Threading Results from Heavy Atom (HA) Pairwise Contact Potentials

L	d_0	n_x	R	R_0	I	$\sigma(I^*)$	D	$\sigma(D^*)$	J	$\sigma(J^*)$	$I/\sigma(I^*)$	$J/\sigma(J^*)$	$\langle Z \rangle$	$I_{\sqrt{n_X}}$
50	4.0	86	2.897	0.001	0.234	0.141	−0.242	0.120	0.477	0.121	1.656	3.933	2.134	2.158
	4.5	141	1.698	0.001	0.222	0.121	−0.215	0.097	0.437	0.105	1.841	4.151	2.618	2.624
	5.0	174	1.644	0.001	0.198	0.111	−0.188	0.088	0.386	0.096	1.783	4.024	2.680	2.595
	5.5	204	1.827	0.001	0.173	0.103	−0.162	0.081	0.335	0.088	1.688	3.810	2.637	2.450
	6.5	283	2.268	0.001	0.126	0.084	−0.117	0.070	0.243	0.069	1.505	3.537	2.520	2.097
100	4.0	179	0.111	0.132	0.227	0.114	−0.243	0.095	0.470	0.087	2.001	5.390	4.009	3.034
	4.5	293	0.059	1.522	0.214	0.096	−0.214	0.077	0.428	0.078	2.220	5.490	4.804	3.647
	5.0	361	0.066	1.878	0.191	0.089	−0.187	0.070	0.377	0.072	2.135	5.217	4.889	3.606
	5.5	421	0.084	2.014	0.166	0.083	−0.161	0.065	0.327	0.067	2.011	4.883	4.790	3.398
	6.5	586	0.152	2.184	0.120	0.068	−0.116	0.055	0.236	0.052	1.770	4.523	4.499	2.894
150	4.0	276	0.005	27.215	0.222	0.101	−0.243	0.084	0.464	0.071	2.204	6.587	5.420	3.666
	4.5	449	0.006	38.150	0.206	0.085	−0.214	0.070	0.419	0.065	2.414	6.457	6.435	4.343
	5.0	553	0.009	38.583	0.183	0.079	−0.186	0.064	0.369	0.061	2.318	6.068	6.532	4.283
	5.5	647	0.010	37.047	0.159	0.073	−0.161	0.059	0.320	0.057	2.168	5.615	6.390	4.028
	6.5	900	0.028	33.310	0.114	0.060	−0.116	0.050	0.230	0.044	1.887	5.236	5.949	3.402
200	4.0	375	0.002	43.738	0.219	0.092	−0.242	0.078	0.461	0.060	2.388	7.679	6.596	4.224
	4.5	608	0.002	43.385	0.200	0.079	−0.213	0.067	0.413	0.058	2.546	7.166	7.790	4.917
	5.0	748	0.003	43.200	0.177	0.073	−0.186	0.061	0.364	0.055	2.427	6.604	7.892	4.838
	5.5	875	0.003	42.113	0.154	0.068	−0.161	0.056	0.314	0.052	2.257	6.084	7.714	4.539
	6.5	1218	0.012	40.743	0.109	0.056	−0.116	0.048	0.225	0.040	1.942	5.706	7.150	3.802
250	4.0	475	0.002	44.372	0.218	0.087	−0.241	0.075	0.459	0.054	2.521	8.550	7.663	4.742
	4.5	767	0.002	42.025	0.197	0.075	−0.213	0.066	0.409	0.054	2.628	7.554	9.018	5.431
	5.0	943	0.003	41.760	0.174	0.070	−0.186	0.061	0.360	0.053	2.484	6.808	9.121	5.327
	5.5	1103	0.003	41.681	0.151	0.065	−0.160	0.057	0.311	0.050	2.310	6.262	8.915	4.994
	6.5	1535	0.008	41.066	0.106	0.054	−0.116	0.048	0.222	0.037	1.968	5.929	8.242	4.148

Table II
Informatic Properties and Threading Results from β-Carbon (CB) Pairwise Contact Potentials

L	d_0	n_x	R	R_0	I	$\sigma(I^*)$	D	$\sigma(D^*)$	J	$\sigma(J^*)$	$I/\sigma(I^*)$	$J/\sigma(J^*)$	$\langle Z \rangle$	$I\sqrt{n_X}$
50	8.5	254	8.179	0.007	0.087	0.070	−0.080	0.065	0.166	0.070	1.229	2.372	1.483	1.363
	9.5	385	8.194	0.001	0.066	0.059	−0.063	0.060	0.130	0.055	1.133	2.376	1.520	1.288
	10.5	569	7.767	0.006	0.055	0.053	−0.054	0.056	0.109	0.046	1.046	2.381	1.598	1.305
	11.5	791	7.301	0.001	0.048	0.048	−0.047	0.052	0.095	0.039	0.996	2.422	1.679	1.333
	12.5	1040	6.843	0.001	0.040	0.043	−0.041	0.049	0.081	0.034	0.928	2.411	1.805	1.285
	13.5	1310	6.946	0.001	0.032	0.039	−0.034	0.045	0.066	0.028	0.820	2.313	1.864	1.140
100	8.5	528	1.540	0.011	0.082	0.056	−0.078	0.052	0.159	0.055	1.466	2.917	3.028	1.862
	9.5	803	1.489	0.036	0.063	0.047	−0.061	0.048	0.125	0.042	1.346	2.973	2.998	1.717
	10.5	1188	1.450	0.032	0.053	0.043	−0.052	0.045	0.105	0.035	1.222	3.004	3.044	1.813
	11.5	1653	1.380	0.099	0.046	0.040	−0.046	0.042	0.092	0.030	1.145	3.050	3.100	1.866
	12.5	2178	1.332	0.624	0.039	0.037	−0.039	0.039	0.078	0.026	1.064	3.064	3.134	1.818
	13.5	2745	1.500	1.348	0.031	0.033	−0.033	0.036	0.064	0.021	0.935	2.984	3.077	1.617
150	8.5	813	0.620	3.095	0.076	0.049	−0.076	0.048	0.152	0.047	1.551	3.228	4.109	2.141
	9.5	1238	0.509	3.277	0.059	0.042	−0.060	0.045	0.119	0.036	1.403	3.338	4.015	2.065
	10.5	1835	0.535	5.045	0.049	0.039	−0.051	0.041	0.100	0.030	1.249	3.384	4.032	2.103
	11.5	2554	0.537	7.852	0.043	0.037	−0.045	0.039	0.088	0.026	1.159	3.437	4.062	2.163
	12.5	3371	0.520	10.301	0.036	0.034	−0.038	0.036	0.075	0.022	1.065	3.456	4.045	2.112
	13.5	4251	0.599	11.262	0.029	0.031	−0.032	0.033	0.061	0.018	0.924	3.399	3.917	1.870
200	8.5	1098	0.401	16.731	0.071	0.045	−0.077	0.048	0.147	0.043	1.569	3.388	4.971	2.329
	9.5	1679	0.269	17.552	0.055	0.039	−0.060	0.044	0.115	0.032	1.399	3.568	4.843	2.248
	10.5	2491	0.311	19.935	0.046	0.038	−0.051	0.041	0.097	0.027	1.221	3.599	4.844	2.280
	11.5	3470	0.340	21.581	0.040	0.035	−0.045	0.038	0.085	0.023	1.126	3.651	4.853	2.334
	12.5	4588	0.312	23.418	0.034	0.033	−0.039	0.036	0.072	0.020	1.021	3.681	4.800	2.277
	13.5	5788	0.342	21.478	0.026	0.030	−0.032	0.033	0.059	0.016	0.871	3.629	4.626	2.002
250	8.5	1381	0.302	27.688	0.066	0.042	−0.078	0.048	0.144	0.042	1.572	3.443	5.742	2.455
	9.5	2117	0.160	29.033	0.052	0.037	−0.061	0.045	0.113	0.030	1.380	3.730	5.590	2.371
	10.5	3147	0.211	30.449	0.043	0.036	−0.052	0.041	0.095	0.025	1.178	3.745	5.572	2.383
	11.5	4385	0.257	31.470	0.037	0.034	−0.046	0.039	0.083	0.022	1.076	3.777	5.563	2.421
	12.5	5809	0.215	33.401	0.031	0.032	−0.040	0.036	0.071	0.018	0.961	3.835	5.480	2.353
	13.5	7336	0.227	31.474	0.024	0.030	−0.033	0.033	0.057	0.015	0.805	3.804	5.272	2.052

from 10.6 to 23.4 cnats for HA, but significantly lower for CB, with a range of 2.4–8.6 cnats. (The unit "cnat" is 1/100 of a "nat," the measure of information arising from the use of the natural logarithm in Eq. (1); if base-2 is used, the unit of information becomes the more familiar "bit.") The difference can be explained by the fact that HA is closer to the physical picture of contacting amino acids within a folded protein environment. The CB definition, on the other hand, approximates contact between two residues by the distance between their C_β atoms, which may not actually be in physical contact. In particular, attractive interactions between residues may not be detected by CB potentials if the distance between C_β atoms happens to be larger than d_0, especially for pairs of residues with bulky side chains. Conversely, for CB potentials with reasonably large d_0, residues with small side chains may be incorrectly classified as contacting pairs. These artifacts (false negatives and false positives) have a negative effect on the pairwise contact mutual information when CB is used.

Mutual information is strongly dependent on the cutoff distance d_0, which defines contact between residue pairs. Maximum information is achieved at the lowest possible d_0, and diminishes at higher d_0, as more residue pairs are considered to be in contact. Low values of d_0 ensure that only actual physical contacts are counted. Liberal definitions of contact (high d_0) dilute the statistics with false positives. As a corollary, mutual information should be expected to asymptotically approach zero as d_0 approaches infinity, the case where *all* residue pairs are taken to be in contact, and no discrimination occurs. If d_0 is allowed to approach the lowest possible contact distances, mutual information might be expected to rise dramatically, as only real contacts will be included in the statistics. However, the number of contacts detected diminishes greatly, which can cause significant statistical instability.

With the comprehensive threading results, the relationships among the informatic quantities I, D, and J can be explored. The absolute value of the divergence D, the average score given by decoy conformations, behaves similarly to mutual information I: higher information values are observed for HA than for CB, as well as a decreasing trend as d_0 is increased. This is confirmed by Figure 3(A),

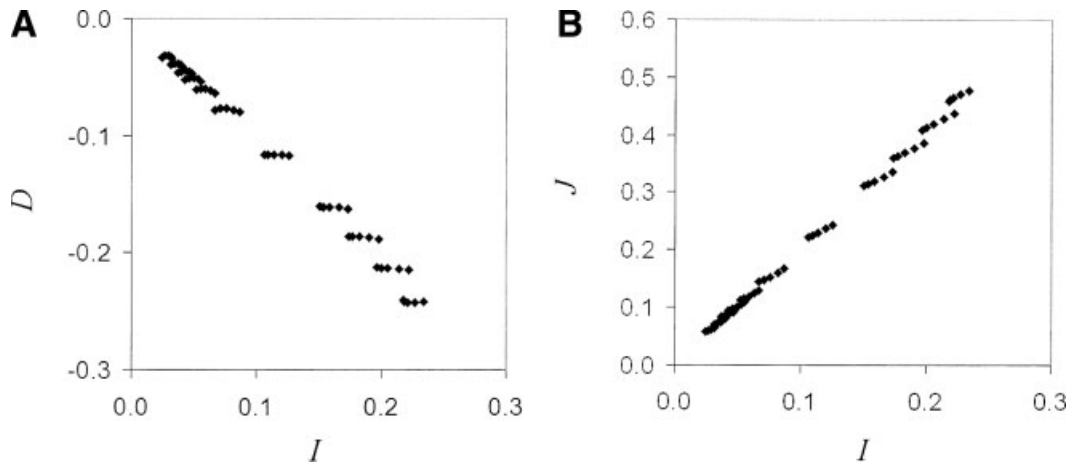

Figure 3

*The relationship between mutual information I and other informatic quantities (**A**) divergence D and (**B**) total divergence J. The threading data for these plots come from all the contact potentials described in Tables I and II, with each data point representing the result of the application of one contact potential in an all-against-all threading. The correlation among these informatic variables is strong (see Table III).*

which shows that the correlation between I and D is significant, implying that contact potentials with higher information content also assign lower (less desirable) scores to decoy conformations. Because I and D sum to J, factors that increase mutual information should also tend to increase discrimination. This is confirmed by Figure 3(B), as well as the correlation measurements summarized in Table III. Thus, total divergence J, the expected difference between the native and non-native score, behaves the same way as I, implying that optimizing potentials for information has the direct effect of increasing score discrimination.

Factors that affect discrimination by contact potentials

In this work, we probe the informatic nature of contact potentials using three specific factors-definition of contact, length of sequence, and cut-off distance-that are known to affect performance. First, our measurements show that discrimination brought about by the contact potential, as measured by J, ranges from 22.2 to 47.7 cnats for HA and from 5.7 to 16.6 cnats for CB. Actual discrimination of the native conformation in a real threading exercise is gauged by $\langle R \rangle$, the mean percentile rank of the native score (with values close to zero signifying superior discrimination), and R_0, the proportion of native conformations assigned the best score. Threading results show that HA potentials perform better than CB at all sequence lengths, based on these two measures.

Second, we find that larger values of sequence length L (at constant d_0) results in better discrimination in both HA and CB potentials. This is true even while both I and J actually decrease at higher L, though only minutely.

Third, data for different cut-off distances of both HA and CB reveal that there appears to be an optimum d_0, which yields the best $\langle R \rangle$ per specified sequence length: for the HA potential at $L = 50$, it is at $d_0 = 5.0$ Å; for $L = 100$, it is at $d_0 = 4.5$ Å; and for $L = 150$, 200, and 250, it is somewhere between $d_0 = 4.0$ and 4.5 Å. Though the measure R_0 selects different optimal potentials at $L = 100$ and 150, it follows the same over-all trend of finding the maxima at decreasing d_0 as L is increased. The same behavior is observed for CB potentials. The informatic functions, however, show consistent increase, across all lengths for both HA and CB, as the distance d_0 is lowered.

We seek a way to reconcile these threading results with information measurements. In particular, we ask how it is possible that potentials with lowered cut-off distances perform poorly despite having significantly higher mutual information and total divergence. The key to this seeming paradox is to recall that these informatic functions are *expectations*. For instance, the quantity $I_L^*(s)$, the native energy of sequence s, is also an estimate of the mutual information $I(a_x a_z)$ of the contact potential, and likewise $J_L^*(s)$ is an estimate of the total divergence $J[(a_x a_z)_N, (a_x a_z)_R]$ of the potential, as in Eq. (14).

Table III

Pearson Product-Moment Correlation Between Informatic Quantities in Threading

	I	D	J
I	1.000		
D	−0.995	1.000	
J	0.999	0.999	1.000

The distribution of $J_L^\star(s)$ around its expected value, the total divergence J, is illustrated in Figure 2. The potential used in this case (HA, $d_0 = 4.5$ Å, applied to $L = 100$) yields a J of 0.428 nats, but as can be seen in the figure, the individual $J_L^\star(s)$ are spread around this average value, each instance corresponding to a particular level of R. Recognizing the logarithmic dependence of R on $J_L^\star(s)$, as demonstrated in the same figure (noting that the y-axis is in the log scale), is crucial to understanding the behavior of $\langle R \rangle$, the mean value given by the spectrum of $J_L^\star(s)$. Thus, we find that the extent of variability of individual $J_L^\star(s)$ around J is a significant factor that can affect $\langle R \rangle$. To illustrate, even at constant J, $\langle R \rangle$ can still vary depending on how *spread out* the individual $J_L^\star(s)$ around the mean value J. That is, the value of $\langle R \rangle$ is increased dramatically if the $J_L^\star(s)$ are clustered tightly around J, while it is diminished if they are more dispersed, despite having the same J.

These observations suggest that aside from the actual value of J, another critical variable to performance is the $\sigma(J^\star)$. Because it is a statistical truism that the number of samples in a typical measurement (in this case, the number of contacts) affects the magnitude of fluctuation of individual $J_L^\star(s)$ cases around the grand average J, then any factor that changes this number will influence $\langle R \rangle$ directly. This relationship can be used to illuminate the patterns in the threading results.

First, the number of contacts is directly proportional to the length of the sequence being threaded, and therefore the variable L must affect threading discrimination. Similarly, d_0 affects the number of contacts defined in a sequence of a given length. The effect of the number of contacts, acting through these two variables, can be seen in Tables I and II. While sequence length L hardly changes the informatic values I, D, and J, the standard deviations $\sigma(I^\star)$, $\sigma(D^\star)$, and $\sigma(J^\star)$ drop significantly as L increases. This is true for both the HA and the CB contact definitions. Threading performance, as measured by $\langle R \rangle$, improves as L increases, illustrating the fact that a high number of score events per sequence is a desirable property.

Next, we examine the action of d_0 on both $\langle R \rangle$ and R_0. We find that the standard deviations $\sigma(I^\star)$, $\sigma(D^\star)$, and $\sigma(J^\star)$ decrease significantly, along with the informatic values, as d_0 is increased. While the variance pattern seen as a function of L is significant, sequence length is not an adjustable variable in real threading applications, since it ostensibly cannot be changed. The quantity d_0, on the other hand, is an independent variable in the construction of contact potentials, and can therefore be optimized. As previously observed, the choice for optimal d_0 is complicated by the fact that those potentials built with d_0's that deliver high I, D, and J, a desirable characteristic, are those that have high variances for their specific quantities, an undesirable characteristic. The optimum d_0 must involve a compromise between the two properties.

One way to combine the action of two opposing properties into one measure is to form a ratio of the (average) quantity and its standard deviation. The resulting "signal-to-noise" ratios—$I/\sigma(I^\star)$, $D/\sigma(D^\star)$, and $J/\sigma(J^\star)$—attempt to calibrate the balance between two quantities that have opposite desirability—in this case, information (or divergence) and fluctuations. To illustrate the behavior of these ratios, we include $I/\sigma(I^\star)$ and $J/\sigma(J^\star)$ in Tables I and II. Indeed, the ratios appear to be a much better indicator of performance, showing peaks at or near the d_0's that show optimum $\langle R \rangle$ and R_0.

Another way to factor in the variability of individual measurements would be to explicitly incorporate the number of contacts, or, more generally, the number of score events. Each sequence–structure alignment, whether correct or incorrect, results in a set of contacts, whose number may differ from conformation to conformation. To simplify the calculation of an "averaged" number of contacts for a threading exercise, we shall pick the number of contacts of the given sequence in its *native* conformation as representative of the range of contacts that the sequence will see as it is threaded through the ensemble of decoy conformations. It will be demonstrated that this choice is not only effective, but also the most practical quantity to use, because it relies only on the native structures, and not on characteristics of the large decoy ensemble.

We can compute the mean number of contacts that a typical sequence (of a certain length) will encounter by employing a given contact potential:

$$n_X = \frac{1}{n_s} \sum\nolimits_i n_i^c \tag{18}$$

where n_i^c is the number of native contacts (or the number of native score events) of sequence i, and the summation runs through all n_s sequences in the data set. Because of the dependency of the standard deviation on sample size, $\sigma \propto 1/\sqrt{n}$, the equivalent signal-to-noise "ratios" look like the following composite quantities: $I\sqrt{n_X}$, $D\sqrt{n_X}$, and $J\sqrt{n_X}$. We shall refer to these as the "information product."

To investigate general relationships between threading success and the informatic ratios, results from comprehensive threading with both HA and CB contact potentials have been combined. Pooling data from these two distinct sets of contact potentials should reveal more general patterns, and should be more indicative of the effect of information on the ability to discriminate native folds.

Foremost, we find that the informatic functions I, D, and J, by themselves, do not show any significant correlation with $\langle R \rangle$ (data not shown). However, when they are normalized by their standard deviations, they show strong correlation when plotted with $\langle R \rangle$ (see Fig. 4). (The correlation values are summarized in Table IV.) In addition, two of the information products, $I\sqrt{n_X}$ and

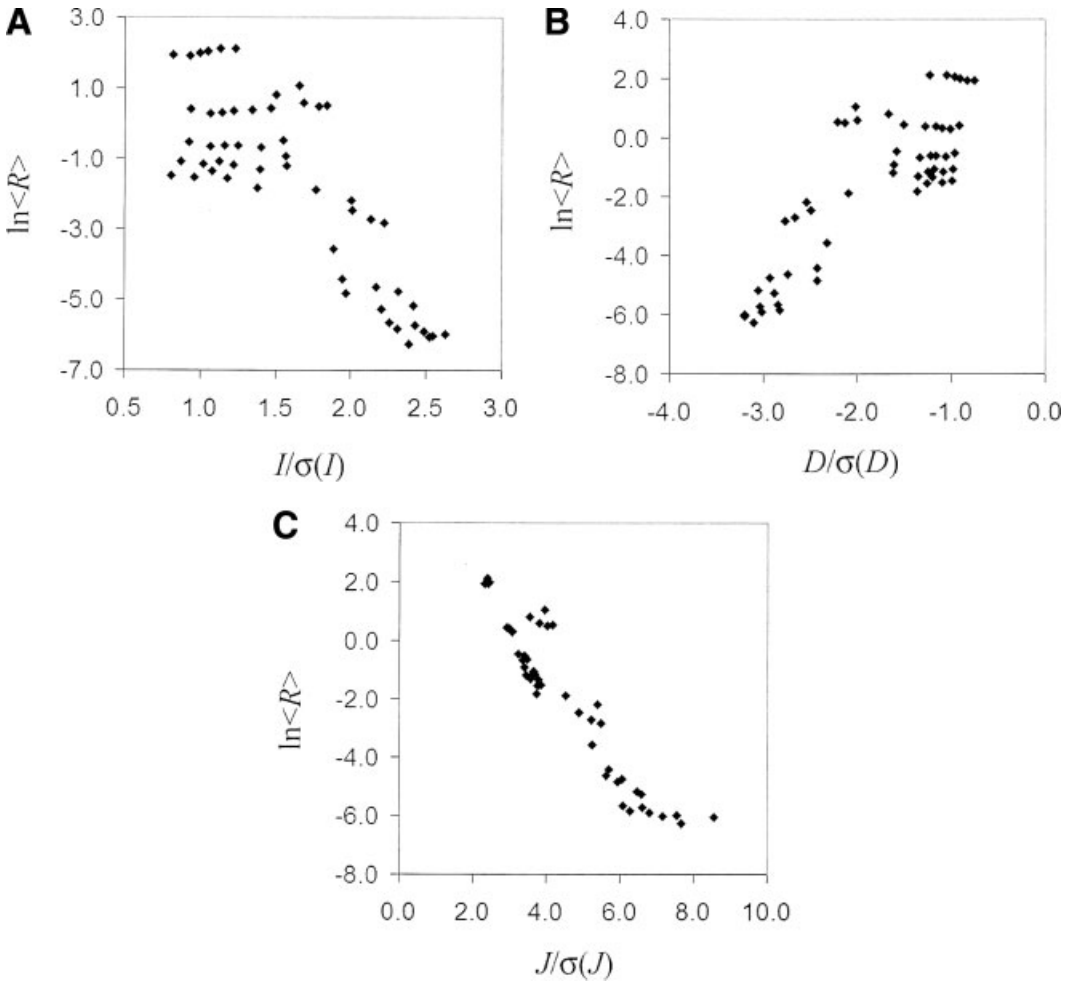

Figure 4

Informatic measures as gauges of the threading behavior of all HA and CB contact potentials examined in this work. The threading data for these plots come from all the contact potentials described in Tables I and II, with each data point representing the result of the application of one contact potential in an all-against-all threading. For each contact potential, at a given length scale, an exhaustive series of threading exercises were implemented, and results are pooled to compute the average values I, D, J, and ⟨R⟩, and the standard deviations σ(I), σ(D), and σ(J). These values are summarized in Tables I and II. The calculation of the informatic ratios ***(A)*** *$I/\sigma(I^*)$,* ***(B)*** *$D/\sigma(D^*)$, and* ***(C)*** *$J/\sigma(J^*)$ are straightforward.*

$J\sqrt{n_X}$, were also plotted with $\langle R \rangle$ (see Fig. 5). Figure 5(B) is partitioned with respect to sequence length L, to demonstrate the effectiveness of $I\sqrt{n_X}$ independent of the variable L. The information product plots show remarkable resemblance with one another, as well as with the performance plot of $J/\sigma(J^*)$.

Taken as a whole, these results indicate that better performance can be achieved by an increase in the information content of the potential, and if possible, an increase in the number of score events per measurement. If fulfilling one condition has the opposite effect on the other, as in the question of d_0, then an optimum is achieved by compromise as embodied in the information product.

We point out that the near-perfect ability to appraise threading performance by the quantity $I\sqrt{n_X}$ has a distinct practical advantage over the other composite quantities considered above (and including the Z-score, discussed below). This quantity can be computed directly from the data set of native structures, without having to do any threading of sequences through decoy conformations. Recall that mutual information I is the average per contact "energy" of native conformations, as expressed in Eq. (8a), while n_X is the average number of contacts of a typical sequence in its native fold. This stands in contrast

Table IV

Pearson Product-Moment Correlation Between Performance and the Composite Informatic Quantities

	$I/\sigma(I^*)$	$D/\sigma(D^*)$	$J/\sigma(J^*)$	$\langle Z \rangle$	$I\sqrt{n_X}$	$J\sqrt{n_X}$
$\langle R \rangle$	−0.813	−0.838	−0.933	−0.939	−0.953	−0.966

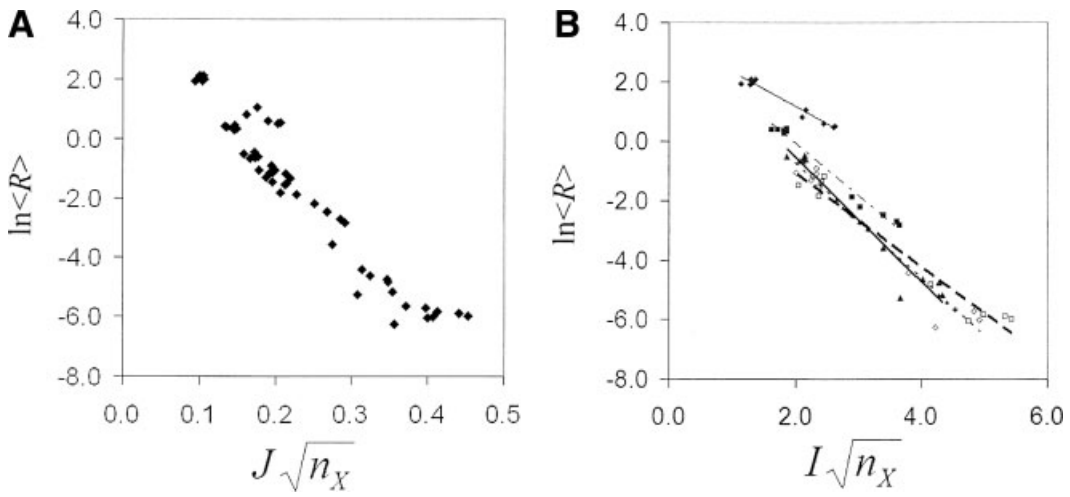

Figure 5

Information product as gauges of the threading behavior of all HA and CB contact potentials examined in this work. The threading data for these plots come from all the contact potentials described in Tables I and II, with each data point representing the result of the application of one contact potential in an all-against-all threading. As in Figure 4, for each contact potential, at a given length scale, an exhaustive series of threading exercises were implemented, and results are pooled to compute relevant informatic quantities and ⟨R⟩. These values are summarized in Tables I and II. The calculation of the information products ***(A)*** *$J\sqrt{n_X}$ and (B) $I\sqrt{n_X}$. In* ***(B)****, the data points have been partitioned by sequence length L. Each line represents a particular L, starting with L = 50 at the top of the plot, down to L = 250 (bold dashed line). Without the regression lines, plot (B) looks almost identical to plot (A).*

with the other composite quantities, which involve the threading of a representative sample of sequences through the decoy ensemble, ostensibly a computationally demanding iterative procedure.

Going back to the issue of threading segments instead of complete sequences, we see from Figure 5(B) that while there are large differences in $\langle R \rangle$, the performance of each segment length follows the same dependence on I and $\sqrt{n_X}$. Small lengths ($L = 50$ and 100) on average may not contain the full amount of information that is necessary to distinguish the native from decoy conformations, but the efficiency by which the incomplete information is extracted from these segments by the potential is still predictive of threading success. The same pattern holds for longer segments ($L = 200$ and 250), which contains a much greater amount of contact folding information.

The relationship among informatic quantities, and the *Z*-score

In the analysis above, the ratio of the informatic quantity and the variance (or standard deviation) has been shown to be indicative of threading performance. The *Z*-score, a common measure of discrimination in statistical methodologies, is analogous to these ratios. The *Z*-score of the native conformation of sequence s is

$$Z(s) = \frac{E_{N|s} - \langle E_{C|s} \rangle}{\sigma_{D^*}(s)} \tag{19a}$$

while in terms of the informatic variables derived thus far,

$$Z(s) = \frac{J_L^*(s)}{\sigma_{D^*}(s)} \tag{19b}$$

where $\sigma_{D^*}(s)$ is the standard deviation of the distribution of scores given by the alignment of sequence s with the ensemble of incorrect conformations $\{C\}$ (i.e., the standard deviation of the specific divergence D^*).

The *Z*-score is an intuitive representation of discrimination for single applications of a folding potential. To characterize the typical action of a contact potential in threading, the mean *Z*-score[25] can be calculated:

$$\langle Z \rangle = \frac{1}{n_{\text{seq}}} \sum_{\substack{\text{all} \\ \text{seqs}}} Z(s) \tag{20}$$

where the summation is over all sequences in the data set, and n_{seq} is the number of chains in the data set. Another way of expressing the mean *Z*-score with informatic functions involves a manipulation of Eq. (19b).[10] Multiplying both sides of this equation by $\sigma_{D^*}(s)$ and taking the mean, we arrive at

$$\langle Z \sigma_{D^*} \rangle = J \tag{21a}$$

The mean of the product on the left hand side of the equation above can be approximated by the product of

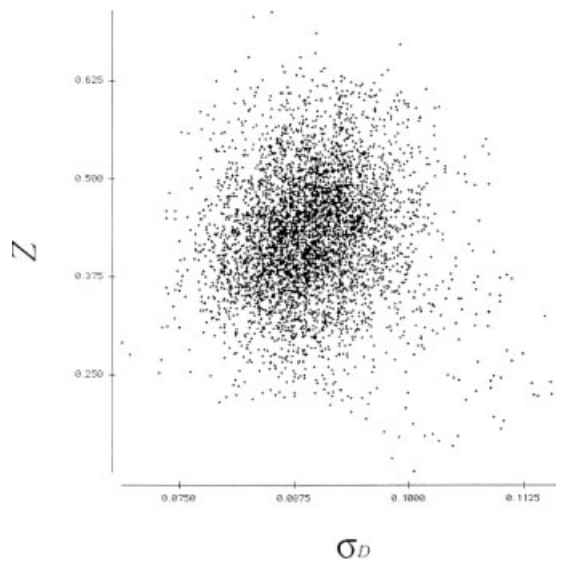

Figure 6

The absence of correlation between the Z-score of the native conformation and the standard deviation of the spectrum of decoy scores, for 5000 randomly selected 100-mer sequences, subjected to a threading procedure with the same contact potential described in Figure 1. The width of the distribution of decoy scores alone is not indicative of the success in identifying the native conformation among an ensemble of decoys. This plot shows that it is possible to deconvolute the mean of the product in Eq. (21a) into the product of two means.

two means if the two variables Z and σ_{D^*} are uncorrelated. Figure 6 confirms this relationship, which allows us to express the mean Z-score as

$$\langle Z \rangle = \frac{J}{\langle \sigma_{D^*} \rangle} \tag{21b}$$

This quantity is operationally similar to the informatic ratios examined in the previous section, but with some key differences. In particular, contrast it with the quantity $J/\sigma(J^*)$, which is the ratio of the expected value of J and the standard deviation of the spectrum of J_L^* values. The denominator $\langle \sigma_{D^*} \rangle$ of the ratio in Eq. (21b) is not the same quantity as $\sigma(D^*)$, used previously in the ratio $D/\sigma(D^*)$. The latter is the standard deviation of the mean scores of many protein sequences mounted onto decoy structures, while the former is the mean value of the standard deviation of the scores given by decoy structures. Nonetheless, $\langle \sigma_{D^*} \rangle$ shows strong correlation with the standard deviations of other informatic quantities (Table V).

$\langle Z \rangle$ values for every contact potential studied in this work are included in Tables I and II. Figure 7 shows the relationship between $\langle Z \rangle$ and threading success across different HA and CB contact potentials examined in this work. There is marked similarity to the performance plot

Table V

Pearson Product-Moment Correlation Between Standard Deviations of Informatic Quantities in Threading.

	$\sigma(I^*)$	$\sigma(D^*)$	$\sigma(J^*)$	$\langle \sigma_{D^*} \rangle$	$\sqrt{n_X}$
$\sigma(I^*)$	1.000				
$\sigma(D^*)$	0.983	1.000			
$\sigma(J^*)$	0.969	0.968	1.000		
$\langle \sigma_{D^*} \rangle$	0.896	0.928	0.940	1.000	
$\sqrt{n_X}$	−0.804	−0.811	−0.854	−0.840	1.000

of $J/\sigma(J^*)$ in Figure 4(C), and $I\sqrt{n_X}$ and $J\sqrt{n_X}$, in Figure 5. Because of the high correlation among $\langle \sigma_{D^*} \rangle$, $\sigma(D^*)$, and $\sqrt{n_X}$, the effectiveness of $\langle Z \rangle$ as a gauge of performance can be traced to the properties of basic informatic functions. But we have not undertaken an informatic dissection of contact potentials simply to demonstrate this connection. By parsing the action of the different parameters of the contact potential, we open up the possibility of direct optimization by parameter manipulation. The $\langle Z \rangle$ measure, on the other hand, while a powerful gauge of performance, has not heretofore been understood directly in terms of the potential parameters, and has remained opaque to direct optimization.

SUMMARY AND CONCLUSION

We seek to understand the nature of long-range pairwise contact information, both in the manner it is

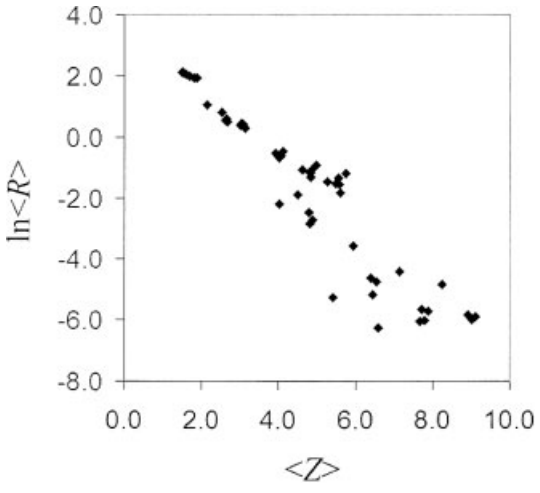

Figure 7

The mean Z-score as indicator of threading success of contact potentials. The quantity $\langle Z \rangle$, built with the same informatic framework as the ratios exhibited in Figure 3, is well correlated with the mean percentile rank of the native score in the spectrum of decoy scores. As in previous figures, the threading data for this plot comes from all the contact potentials described in Tables I and II, with each data point representing the result of the application of one contact potential in an all-against-all threading.

encoded in sequence and its ability to discriminate among native-like conformations, with the primary interest in optimizing performance. In this work, we use basic information-theoretic ideas to carry out two critical analyses: first, to quantify the information stored in pairwise contacts, and then to understand the behavior of contact potentials in discriminating native from non-native conformations as a function of this information.

In an effort to properly measure the relationship between information and performance, we look at the behavior of the contact potentials developed here across a range of protein folds and sequences of various lengths. Given the native conformations of a large set of non-redundant protein sequences, the best measurements that can be derived are those that result from exhaustive threading of *all* sequence segments in the data set of given length L through *all* possible folded conformations. We demonstrate here that informatic quantities that characterize a given potential arise naturally from such an all-against-all threading operation.

In particular, we establish the connection between information-theoretic quantities and the contact "energies" or scores used in protein structure prediction. We have established the following equivalences:

1. The expected "energy" or score of a correct sequence–structure alignment and mutual information I;
2. The expected score of incorrect alignments in a threading exercise and divergence D; and
3. The expected gap between the native score and the mean score of the ensemble of incorrect conformations, and total divergence J.

We find that pairwise contacts between residues in a protein chain bear a moderate amount of folding information. Our data shows that contact potentials designed to summarize such interactions can reach levels of 20 cnats or more. In the framework of threading, the total divergence J is the operative quantity of discrimination. It is natural to assume that the information content *and* total divergence of contact potentials are directly responsible for their consistent effectiveness in threading applications, as shown in this work and in numerous previous studies.

For contact potentials, we observe that I is sufficiently correlated with D, and therefore an increase in I results in an increase in J. This is the reason why knowledge of mutual information I is enough to characterize J, the discrimination gap. The relationships among these three informatic quantities I, D, and J make intuitive sense if one views the search for the correct fold probabilistically, as follows. A potential with an increased mutual information I means that *higher* probabilities are being assigned to the correct conformations, which implies that it also assigns *lower* probabilities to incorrect conformations. Thus, decoys (incorrect conformations) will receive lower scores on average, which is reflected in a lowered $-D$. Thus, the effect of increasing mutual information in statistical potentials has the effect of making the informational landscape easier to search, since there will be steeper and taller peaks around the native conformation, and lower valleys for the rest of the landscape.

The definition of pair contact is critical to the information content, and therefore to the performance of contact potentials in fold recognition. Defining contact by using the shortest distance between the heavy atoms that compose the residue pair (HA) generates more mutual information compared to using the distance between C_β atoms (CB). Indeed, threading results confirm that HA potentials are able to discriminate correct from incorrect conformation more effectively. Similarly, the contact cut-off distance d_0 has a critical effect on the performance of contact potentials. This parameter is adjustable. We find that for any given contact potential, there exists an optimum d_0, which maximizes its performance. We also find that mutual information in contact potentials increases as d_0 is decreased.

Potentials that have been designed to have high information content should be superior in identifying the correct fold. The discriminatory power, as measured by $\langle R \rangle$, the mean percentile rank of the native conformation, is correlated with the informatic values normalized by the width of their ranges. These composite quantities, such as the information ratio $J/\sigma(J^\star)$ and the information product $I\sqrt{n_X}$, measure the combined fitness of two quantities that have opposite desirability—in this case, information (or divergence) and the number of score events. The dependence of performance on such quantities is clearly illustrated in the optimization of contact potentials with respect to d_0. Contact potentials that use low d_0 values do contain higher amounts of mutual information (and therefore, total divergence), but may be too restrictive to include enough contacts to generate a stable score value per sequence. Because of this significant fluctuation, the effectiveness of such potentials is diminished.

We observe that the information product $I\sqrt{n_X}$ is a superior indicator of threading success, as measured by $\langle R \rangle$, the mean percentile rank of native conformations. We find two key conditions for better discrimination of the native conformation by a statistical potential: high mean native scores and a high number of score events per typical measurement. To satisfy the first condition, higher mean native scores may be obtained from an increase in the information content I of the potential. This is accomplished by using information maximization methods developed in previous work.[10] As to the second condition, we find that parameters that involve an increase in score events can serve to improve potential discrimination. There may be parameters, which bring about opposite desirability (in terms of these two conditions) when varied. For instance, decreasing d_0 raises the

mutual information, but lowers the number of detected contacts. In such cases, a compromise value of the parameter exists that optimizes performance.

Analogous to these composite quantities is the more common *Z*-score, a statistical quantity used to measure the relative position of the native score of a sequence in the spectrum of scores given by the universe of conformations. It is not surprising to find operational similarities, in fold recognition and threading, of $\langle Z \rangle$ with the composite quantities derived here, including the information product. However, the success of $I\sqrt{n_X}$ in gauging performance offers a clear advantage. Recall that this quantity can be computed directly from the data set of native folds. Operationally, this means that threading success may be optimized by selecting potentials that maximize this quantity, without any need for actual threading with decoy ensembles (i.e., $\langle Z \rangle$ optimization[25]). This presents an opportunity to optimize potentials, with knowledge only of the native scores, by direct information maximization.

Other potentials that describe important intramolecular interactions can be viewed through this result. In particular, if varying any of the adjustable parameters does not change n_X, then the performance of the potential can be gauged solely using mutual information *I*. For instance, distance-dependent pairwise side-chain potentials,[34] an analogue of contact potentials, generally have the same number of score events per given length of the chain if cut-off distances are not employed, and all possible pairs of residues at any distance are assigned scores. We are working to confirm the correlation between $\langle R \rangle$ and *I* in distance-dependent pair potentials. Previous work on local-sequence backbone potentials,[14] which have constant n_X, demonstrate clearly that knowledge of *I* alone can indicate how well the potential does in fold recognition, and that any factor that increases mutual information should work to benefit performance.

Moreover, from a statistical viewpoint, as the number of score events increases, one should expect the stability of the scores of individual sequences (with respect to the mean score *I*) to increase as well. This situation describes the application of comprehensive potentials, which take into account various interactions known to be critical to protein stability (i.e., local sequence–backbone interaction, solvation, and general nonlocal/contact effects). Specifically, such potentials ought to have a substantially high number of score events, making the total information content especially relevant. This suggests that if over-all information is maximized in building those potentials, their optimal performance in fold recognition may well be achieved.

Lastly, we note that this work articulates an important role for information-theoretic techniques in the probabilistic analysis of statistical potentials for folding and fold recognition. By demonstrating the equivalence of the quasi-chemically approximated "energies" to fundamental informatic quantities, we bring statistical potentials a step closer to bioinformatics, while freeing them from biophysical approximations. Statistical potentials are a tool, which can be used to detect and utilize information that is available in empirical data, whether or not they are easily related to real physical forces.

REFERENCES

1. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
2. 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.
3. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins: Struct Funct Genet 2000;38:3–16.
4. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins: Struct Funct Bioinform 2005;59:49–57.
5. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein-sequence through the folding motif. Proteins: Struct Funct Genet 1993;16:92–112.
7. 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.
8. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom–atom contact scoring. Proc Natl Acad Sci USA 2003;100:3215–3220.
9. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
10. Solis AD, Rackovsky S. Improvement of statistical potentials and threading score functions using information maximization. Proteins: Struct Funct Bioinform 2006;62:892–908.
11. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogerts RG, Haussler D. Information-theoretic dissection of pairwise contact potentials. Proteins: Struct Funct Genet 2002;49:7–14.
12. Crooks GE, Wolfe J, Brenner SE. Measurements of protein sequence–structure correlations. Proteins: Struct Funct Bioinform 2004;57:804–810.
13. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins: Struct Funct Genet 2000;38:149–164.
14. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins: Struct Funct Genet 2002;48:463–486.
15. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
16. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996; 93:11628–11633.
17. Ben-Naim A. Statistical potentials extracted from protein structures: are these meaningful potentials? J Chem Phys 1997;107:3698–3706.
18. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
19. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
20. Samudrala R, Moult J. Determinants of side chain conformational preferences in protein structures. Protein Eng 1998;11:991–998.

21. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
22. Wang K, Fain B, Levitt M, Samudrala R. Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol 2004;4:8.
23. Dehouck Y, Gillis D, Rooman M. A New Generation of statistical potentials for proteins. Biophys J 2006;90:4010–4017.
24. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
25. 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.
26. Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1991.
27. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995; 4:2107–2117.
28. Godzik A. Knowledge-based potentials for protein folding: what can we learn from known protein structures? Structure 1996;4:363–366.
29. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
30. Zhang C, Liu S, Zhou H, Zhou Y. The dependence of all-atom statistical potentials on structural training database. Biophys J 2004;86:3349–3358.
31. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676–688.
32. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
33. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
34. Sippl M. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.