Published in final edited form as:

Proteins. 2010 May 1; 78(6): 1382–1397. doi:10.1002/prot.22652.

Information-Theoretic Analysis of the Reference State in Contact Potentials used for Protein Structure Prediction

Armando D Solis^{1,2,*} and Shalom R Rackovsky¹

Shalom R Rackovsky: shalom.rackovsky@mssm.edu

¹ Department of Pharmacology and Systems Therapeutics, Box 1603, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, NY 10029, U.S.A

² Department of Biological Sciences, New York City College of Technology, City University of New York, 300 Jay Street, Brooklyn, NY 11201

Abstract

Using information-theoretic concepts, we examine the role of the reference state, a crucial component of empirical potential functions, in protein fold recognition. We derive an information-based connection between the probability distribution functions of the reference state and those that characterize the decoy set used in threading. In examining commonly used contact reference states, we find that the quasi-chemical (QC) approximation is informatically superior to other variant models designed to include characteristics of real protein chains, such as finite length and variable amino acid composition from protein to protein. We observe that in these variant models, the total divergence, the operative function that quantifies discrimination, decreases along with threading performance. We find that any amount of nativeness encoded in the reference state model does not significantly improve threading performance. A promising avenue for the development of better potentials is suggested by our information-theoretic analysis of the action of contact potentials on *individual* protein sequences. Our results show that contact potentials perform better when the compositional properties of the data set used to derive the score function probabilities are similar to the properties of the sequence of interest. Results also suggest to use only sequences of similar composition in deriving contact potentials, to tailor the contact potential specifically for a test sequence.

1. INTRODUCTION

The prediction of protein structure requires conformational-energy-based score functions that can correctly pick the native conformation out of a large number of incorrect folds. In order to properly evaluate the nativeness of the interactions in a given conformation, its conformational energy is measured relative to a so-called reference state, a hypothetical “random” state where those interactions are absent. A common empirical approach is to construct this energy using the Boltzmann formalism,^{1, 2} quantifying it as a log-odds ratio of two probabilities: the probability of finding the query sequence in a given conformation under native conditions, and the probability of its occurrence in the reference state. The former, the so-called “observed” probability, is usually estimated from a statistical survey of experimentally solved protein conformations. The estimation of the latter, the “expected” or reference probability, has proven to be a difficult task, because this state is inaccessible by direct experimental observation. Computational modelling of the hypothetical “random” state is not straightforward either. This uncertainty has led to the development of a number

*To whom correspondence should be addressed: asolis@citytech.cuny.edu, phone: (212) 241-4868, fax: (212) 996-7214.

of reference state models, giving rise to the variety of empirical energy functions found in the literature.³

Empirical energy or score functions have, in recent years, performed increasingly well under stringent computational assessment. This is because such functions, however modelled, are statistical in nature.⁴ They can be taken as a quantitative summary of the sequence-dependent structural information found in native folds. In previous work, we have applied concepts of information theory to quantify such structural information,^{5, 6} and have formulated information-based methods to make statistical potentials more effective in structure prediction.^{7, 8} In particular, we have demonstrated that the way these sequence-dependent probabilities are defined affects the amount of information that can be extracted from empirical data. Consequently, we have developed methods to optimize descriptions of sequence and conformation to maximize performance in structure prediction. In the present work, we use the same information-theoretic tools to explore the reference state problem. The advantage of an information-based approach is that it allows us to bypass complex biophysical considerations, and examine directly the statistical and informatic properties of score functions.

We use our information-based methodology to examine the effect of the choice of reference state model on the effectiveness of potentials involving contacts between side chains of residues in the protein chain. Contact potentials are used widely because of their respectable performance in fold recognition, relative simplicity, and undemanding parameterization.^{9–11} In reality, one can choose any reference state from which to measure energies or scores. Though its precise meaning is open to interpretation, the concept of “expected” probability can provide initial guidance. Early models of contact energy¹² assumed that the expected probability of contact between any two amino acids in a folded protein should be proportional to their mole fractions. This model, the so-called quasichemical approximation, has proven to be effective in parameterizing contact energy, despite the fact that it neglects correlations that arise from the connectivity of the chain. Improvements to the reference state to account for chain connectivity and other properties of folded proteins have been made,¹⁰ but it has been shown that many of these improved models can be easily reduced to the simpler quasichemical reference state,¹³ and provide only modest performance improvement in fold recognition. In the meantime, other contact energy reference states have been advanced in the literature as alternatives to the quasichemical approximation, using different models for the “expected” probability.^{10, 14–17}

Despite growing empirical evidence that variants of the quasichemical approximation work equally well, there is still no clear consensus on how to derive the best-performing reference state. (We should note that there is also a parallel set of investigations for reference states for distance-dependent energy functions,^{18–20} which we hope to address in future work.) In this work, we revisit the contact reference state problem using information-theoretic tools we have developed in previous work. We have found that a key determinant of the correct discrimination of native folds amidst an ensemble of incorrect or “decoy” conformations is the total divergence, an information-theoretic entity that quantifies the distance between the score of the correct structure and the mean score of the decoys.⁸ Here, we demonstrate how the definition of “expected” probability affects the total divergence of contact potentials, and evaluate the impact of the definitions on their effectiveness in actual threading. In the course of our investigation, we discover a connection between the properties of the data set from which a potential is derived, and the properties of the particular query sequence on which the potential will be used. In effect, we formulate a basis for query-specific contact potentials, which have been shown to improve performance. Our goal is to understand how the choice of reference state affects the quantity of information that can be extracted from empirical data, in order to maximize data use in structure prediction efforts.

2. THEORY AND ANALYSIS

2.1. Divergence and total divergence

We begin by outlining the information-theoretic tools that will be used in the analysis. The information-theoretic divergence

$$D(X \| Y) = \sum_{\text{all states}} p(x) \ln \frac{p(x)}{p(y)} \quad (1)$$

is used routinely to measure the distance between discrete probability distributions describing random variables X and Y . Strictly speaking, this is not true a distance, because it is not symmetric, and does not satisfy the triangle inequality. A related measure is the total divergence

$$\begin{aligned} J(X, Y) &= D(X \| Y) + D(Y \| X) \\ &= \sum p(x) \log \frac{p(x)}{p(y)} + \sum p(y) \log \frac{p(y)}{p(x)} \end{aligned} \quad (2)$$

which, unlike D , is symmetric. A useful property of divergence is that

$$D(X \| Y) \geq 0 \quad (3)$$

with equality iff $p(x) = p(y)$ across all states.²¹ Eq. (3) can be used to derive another inequality, as follows:

$$\begin{aligned} \sum p(x) \log \frac{p(x)}{p(x')} &\geq 0 \\ \sum p(x) \left[\log \frac{p(x)}{p(y)} - \log \frac{p(x)}{p(y')} \right] &\geq 0, \text{ and therefore} \\ \sum p(x) \log \frac{p(x)}{p(y)} &\geq \sum p(x) \log \frac{p(x)}{p(y')} \end{aligned} \quad (4)$$

with equality iff $\{p(x)\} = \{p(x')\}$ for all states. This inequality will prove useful at a number of points in this work. One immediate implication is in estimating probabilities, critical in building empirical potentials. The inequality indicates diminished divergence if the true underlying probability distribution $p(x)$ is poorly approximated by $p(x')$, i.e. $\{p(x)\} \neq \{p(x')\}$. Therefore, since the total divergence J of a potential is indicative of performance,^{7, 8} accurate approximations of probabilities from empirical data is critical.

2.2. Score Function and Total Divergence

The informatic quantities described above have been used to model sequence-structure alignment or threading. We extend previous results^{7, 8} here to examine the importance of the reference state to the effectiveness of the resulting score function. Typically, the score of an alignment of a query sequence s and a test conformation c is an additive potential, built from empirical data:

$$E_q(c|s) = \sum_i^m e_q(c_i|s) \quad (5a)$$

where $e_q(c_i|s)$ is a log-odds score function:

$$e_q(c_i|s) = \log \frac{p^{\text{obs}}(c_i|s, q)}{p^{\text{exp}}(c_i|s)} \quad (5b)$$

The numerator $p^{\text{obs}}(c_i|s, q)$ is the observed frequency of the particular (c_i, s) alignment in the database of native folds, a straightforward computation. The denominator $p^{\text{exp}}(c_i|s)$ is the reference state of the score function, which can be imagined as the expected frequency of the same sequence-conformation pair when the interaction of interest, q , is assumed to be absent. (We note that energy-type scores are the negative of the expression in Eq. (5b); however, we choose to define the scores as positive, consistent with information-theoretic usage.) The summation used in the total score $E_q(c|s)$ covers all n_x score instances in the protein chain. The per-interaction score can be defined by:

$$E'_q(c|s) = \frac{1}{n_x} E_q(c|s) = \frac{1}{n_x} \sum_i^m e_q(c_i|s) \quad (6)$$

Various interactions q have been quantitated by this scoring scheme.^{4, 6, 8, 22-26} If the interaction q is significant in protein stability, the score function $E'_q(c|s)$ can be effective in evaluating the fitness of any given sequence-structure alignment.

The gapless threading procedure, a good model for structure prediction, involves comparison of the score of the native (correct) conformation with the spectrum of scores given by an extensive ensemble of decoy (incorrect) structures.²⁷ Correct detection occurs when the score of the native conformation is highest (or the corresponding energy is lowest). Expected behavior of a threading potential can be evaluated by repeated use of the potential in a battery of fold recognition tests. In previous work, we have shown that one quantity used to evaluate discrimination success is related to well-known information-theoretic properties of the scoring function.^{7, 8} This quantity is the gap between native and decoy scores. For a typical sequence s , this gap is

$$J_q(c, s) = E'_q(c^N|s) - \frac{1}{n_d} \sum_j E'_q(c^j|s) \quad (7)$$

where c^N refers to the native conformation, and the summation runs through the ensemble of n_d decoy conformations. Upon repeated applications of the score function to threading of a representative set of n_s sequences, an average gap is derived:

$$\begin{aligned} J_q(C, S) &= \frac{1}{n_s} \sum_k \left[E'_q(c^N|s_k) - \frac{1}{n_d} \sum_j E'_q(c^j|s_k) \right] \\ &= \frac{1}{n_s} \sum_k E'_q(c^N|s_k) - \frac{1}{n_s} \sum_k \frac{1}{n_d} \sum_j E'_q(c^j|s_k) \end{aligned} \quad (8a)$$

or, in terms of the basic scoring function $e_q(c|s)$,

$$J_q(C, S) = \frac{1}{n_s} \sum_k \frac{1}{n_x} \sum_i e_q(c_i^N | s_k) - \frac{1}{n_s} \sum_k \frac{1}{n_d} \sum_j \frac{1}{n_x} \sum_i e_q(c_i^j | s_k) \quad (8b)$$

The first term of the right hand side represents the average per-interaction score given by a sequence in its native conformation, which we have shown in previous work to be equal to mutual information between sequence and conformation.^{7,8} The second term is the expected per-interaction score given by a given sequence mounted onto a typical decoy conformation. To simplify the equation further, we recognize that in repeated threading, the total numbers of sequences n_s and decoy conformations n_d are constant. Depending on the interaction q being studied, the number of scoring instances n_x may or may not vary with the structural details of each decoy. If we assume that it is constant across the decoy ensemble, we find that:

$$J_q(C, S) = \sum_k \sum_i \frac{1}{n_s n_x} e_q(c_i^N | s_k) - \sum_k \sum_i \sum_j \frac{1}{n_s n_d n_x} e_q(c_i^j | s_k) \quad (9)$$

The summations above run through each instance of (c_i, s_k) alignment in the native and decoy ensembles. To check the validity of the simplification of constant n_x We computed the second term of the righthand side of Eqs. 8b and 9 for our all-against-all threading exercise (to be described in Section 3), and found that both quantities agree to at least 5 significant figures.

Another way to express this equation is to count the instances of each *unique* (c, s) alignment, and then recast it as summations through all unique pairs. For instance, if the (c_r, s_t) pair occurs $n^N(c_r, s_t)$ times in the native set and $n^D(c_r, s_t)$ times in the threading of sequence s_t through the decoy ensemble, then the expectation of the score gap can be rewritten as follows:

$$J_q(C, S) = \sum_{r,t} \frac{n^N(c_r, s_t)}{n_s n_x} e_q(c_r | s_t) - \sum_{r,t} \frac{n^D(c_r, s_t)}{n_s n_d n_x} e_q(c_r | s_t) \quad (10a)$$

where the summation runs through all unique sequence-conformation pairs. The frequency ratios can be represented by more familiar notation:

$$J_q(C, S) = \sum_{r,t} p^N(c_r, s_t) e_q(c_r | s_t) - \sum_{r,t} p^D(c_r, s_t) e_q(c_r | s_t) \quad (10b)$$

while, using the score function (Eq. (5b)), we have:

$$J_q(C, S) = \sum_{r,t} p^N(c_r, s_t) \log \frac{p^{\text{obs}}(c_r | s_t, q)}{p^{\text{exp}}(c_r | s_t)} - \sum_{r,t} p^D(c_r, s_t) \log \frac{p^{\text{obs}}(c_r | s_t, q)}{p^{\text{exp}}(c_r | s_t)} \quad (10c)$$

This equation can be converted into a more familiar information-theoretic formulation, by multiplying the numerator and denominator of the score function by $p(s_k)$, and reversing the sign of the decoy term:

$$J_q(C, S) = \sum_{r,t} p^N(c_r, s_t) \log \frac{p^{\text{obs}}(c_r, s_t|q)}{p^{\text{exp}}(c_r, s_t)} + \sum_{r,t} p^D(c_r, s_t) \log \frac{p^{\text{exp}}(c_r, s_t)}{p^{\text{obs}}(c_r, s_t|q)} \quad (11)$$

In summary, the expected gap between the native score and the mean score of decoy conformations, represented by $J_q(C, S)$, is the sum of two means: the first term is the mean native score, while the second term is the (negative of the) mean of the scores given by the ensemble of decoys. (We shall henceforth refer to these two terms as the native term and the decoy term, respectively.)

Comparing to Eq. (2), it is easy to recognize that $J_q(C, S)$ has the form of a total divergence, but only under two specific conditions:

$$p^{\text{obs}}(c_r, s_t|q) = p^N(c_r, s_t) \quad (12)$$

$$\text{and } p^{\text{exp}}(c_r, s_t) = p^D(c_r, s_t) \quad (13)$$

The quantities on the left hand side are components of the score function $e_q(c|s)$, freely adjustable and limited only by the definitions of “observed” and “expected” probabilities. In contrast, the quantities on the right hand side are empirical probabilities that can be derived directly from data sets of experimental (native) and decoy structures. Enforcing these two equalities has the practical consequence of constructing score functions directly from empirical data, or more accurately, from the kind of data (skewed or otherwise) to which the score function will be applied. These conceptual bridges clarify the concepts of the reference state and its “expected” probabilities, which have been difficult to define theoretically.

2.3. Maximization of Total Divergence

We now examine issues relating to the two pairs of probability functions more closely. The first, expressed in Eq. (12), is an equality widely accepted in computational biology, but is only valid under a strict condition—that the expectation of the empirical probabilities characterizing the native state may be assumed to be identical to the probabilities observed from empirical data only when the data set is sufficiently representative of the diversity of protein sequences and structures.

The second condition (Eq. (13)), defining the nature of the reference state, is of primary interest here. We shall explore the consequences of a choice of reference state and gauge the performance of the resulting score functions in threading. The other important characteristic of score functions, the variance of scores, is also examined.

From Eq. (4), it can be seen that the two equalities maximize the native and decoy terms individually:

$$\sum_{r,t} p^N(c_r, s_t) \log \frac{p^N(c_r, s_t)}{p^D(c_r, s_t)} \geq \sum_{r,t} p^N(c_r, s_t) \log \frac{p'(c_r, s_t)}{p^D(c_r, s_t)} \quad (14)$$

for the native term, and

$$\sum_{r,t} p^D(c_r, s_t) \log \frac{p^D(c_r, s_t)}{p^N(c_r, s_t)} \geq \sum_{r,t} p^D(c_r, s_t) \log \frac{p'(c_r, s_t)}{p^N(c_r, s_t)} \quad (15)$$

for the decoy term. The information-based optimization implemented previously⁶⁻⁸ employs the strategy of maximizing the native term. In those studies, we found that factors that increase mutual information (the left hand side of the inequality in Eq. (14)) also increase $J_q(C, S)$, improving threading performance. We extend that work here, by addressing issues surrounding the maximization of the decoy term.

The prescription given by Eqs. (14) and (15) when applied to the score function, however, does not guarantee maximization of $J_q(C, S)$. This is because these inequalities do not deal explicitly with the denominator of the log-ratio. In particular, while choosing $p^D(c_r, s_t)$ over $p'(c_r, s_t)$ for the decoy term is desirable according to Eq. (15), one cannot say anything concrete about the relative magnitudes of two possible native terms

$\sum_{r,t} p^N(c_r, s_t) \log \frac{p^N(c_r, s_t)}{p^D(c_r, s_t)}$ and $\sum_{r,t} p^N(c_r, s_t) \log \frac{p^N(c_r, s_t)}{p'(c_r, s_t)}$ that result from decoy term optimization. The choice of reference state $p^D(c_r, s_t)$ which maximizes the r t decoy term in $J_q(C, S)$ in Eq. (15) may not be consistent with the maximization of the native score term. There may exist another reference state $p'(c_r, s_t)$ which, while r t diminishing the value of the decoy score term, increases the native score term enough to offset it, resulting in a higher sum $J_q(C, S)$. Nevertheless, the fact remains that adopting the equalities of Eqs. (12) and (13) provides at least the local optimization of the two components of $J_q(C, S)$.

3. RESULTS AND DISCUSSION

We explore issues regarding the reference state correctly by using pairwise contact potentials. Details of the comprehensive threading procedure can be found elsewhere.⁸ Briefly, using a set of representative X-ray structures of protein chains, we model the gapless threading exercise by designing an all-against-all test. This procedure involves finding the score-rank of the native conformation of *every* sequence chain in the data set with respect to the ensemble of incorrect conformations provided by the same data set. In recent work,⁸ we demonstrated that measurements from comprehensive threading tests correspond to the components of the total divergence equation $J_q(C, S)$: namely, that the mean score of native conformations is equal to mutual information, and the mean score of incorrect conformations is equal to the directed divergence.

We rewrite the equations derived above in terms of the contact potential. The score function is

$$e_c(c|ab) = \log \frac{p^{\text{obs}}(c|ab)}{p^{\text{exp}}(c|ab)} \quad (16)$$

while the total divergence, or mean score gap, is:

$$J_q(C, S) = \sum_{ab} p^N(c|ab) \log \frac{p^{\text{obs}}(c|ab)}{p^{\text{exp}}(c|ab)} + \sum_{ab} p^D(c|ab) \log \frac{p^{\text{exp}}(c|ab)}{p^{\text{obs}}(c|ab)} \quad (17)$$

Lastly, the maximization condition for the decoy term is:

$$p^{\text{exp}}(c|ab)=p^D(c|ab) \quad (18)$$

In these equations, $p(c|ab)$ refers to the probability of contact between amino acid pair ab .

To define the contact potential, amino acid pairs are represented by their beta-carbons (alpha carbon for glycine). Contact occurs between two side chains if their representative beta (or alpha) carbon atoms are within 9.5Å. All-against-all threading was implemented with 150-mer sequences, mounted onto all continuous 150-mer conformations in the database. With a data set of high-resolution X-ray structures of 1036 proteins, made up of 210,995 residues, a total of 58,034 150-mer sequences were aligned with each one of the same number of conformations, and their scores tallied.

3.1. Perturbing the Reference State Probabilities

As an initial exercise, we have designed a simple experiment to track the behavior of $J_q(C, S)$ upon random perturbation of the reference state probabilities $p^{\text{exp}}(c|ab)$ in the score function. For the base reference state, we set the expected probability of contact between two amino acids a and b to equal the product of their mole fractions, as they occur in the data set:

$$p_{\text{QC}}^{\text{exp}}(c|ab)=k\chi_a\chi_b=k\left(\frac{\sum_i n_i(a)}{\sum_i N_i}\right)\left(\frac{\sum_i n_i(b)}{\sum_i N_i}\right) \quad (19)$$

where $n_i(a)$ is the number of a amino acids in protein chain i , $k=2$ if $a \neq b$ and 1 otherwise, N_i is the total number of residues in protein chain i , and the summation covers all protein chains in the data set. This assumption, called the quasi-chemical approximation (which we will henceforth call QC), is commonly employed in contact potentials, and has been shown to work effectively in gapless threading.¹³

The observed probability distribution component, $p^{\text{obs}}(c|ab)$, is derived from frequency counts of native contacts in the data set. The present simulation entails random perturbations of $p^{\text{exp}}(c|ab)$ from the initial quasi-chemical reference, while keeping $p^{\text{obs}}(c|ab)$ constant, to create entirely new score functions. Perturbations of varying degrees are made to the probability distribution, in order to explore a wide range of reference states relative to QC. The effectiveness of each newly generated score function is evaluated using the all-against-all threading.

We generated 400 unique reference state distributions, yielding changes in the mean gap score $J_q(C, S)$ and $\sigma(D^*)$, the standard deviation of the mean decoy score. Plots of these, as well as of the threading results, are shown in Figure 1. We measure the distance between the starting QC model and any other reference state using the total divergence equation (Eq. (2)), with the form:

$$J_q(p_{\text{QC}}^{\text{exp}}, p_{\delta}^{\text{exp}})=\sum_{ab} p_{\text{QC}}^{\text{exp}}(c|ab) \log \frac{p_{\text{QC}}^{\text{exp}}(c|ab)}{p_{\delta}^{\text{exp}}(c|ab)} + \sum_{ab} p_{\delta}^{\text{exp}}(c|ab) \log \frac{p_{\delta}^{\text{exp}}(c|ab)}{p_{\text{QC}}^{\text{exp}}(c|ab)} \quad (20)$$

where $p_{\delta}^{\text{exp}}(c|ab)$ is the perturbed probability. We find that a modest number of the reference states (5%) actually yield a larger $J_q(C, S)$, though by only a small margin (see Figure 1A). However, as the distance of the perturbed reference state from the quasi-chemical reference

state $J_q(p_{QC}^{exp}, p_{\delta}^{exp})$ increases, it is more likely to generate a smaller gap. A similar pattern can be observed in $\sigma(D^*)$ (Figure 1B), which is a component of Z-score optimization.²⁸ About 33% of the perturbed models, mostly in the vicinity of the QC, yield marginally lower decoy score variance. However, higher score variances are more likely as the distance from QC increases.

Effectiveness of a score function can be measured by the relative rank of the native conformation in relation to the decoy ensemble. A rank $r(s)$ of the native score of 1 signifies that the native score is the best over-all. The mean percentile rank,

$$\langle r \rangle = \frac{1}{n_{L-mer}} \sum_{\text{all seqs}} r(s) \quad (21)$$

computed from all-against-all threading, is the most stringent gauge of performance. In this set of 400 perturbed reference states, all except 20 have a higher $\langle r \rangle$ than the quasi-chemical reference state (which corresponds to poorer discrimination of the native conformation). The dependence of $\langle r \rangle$ on the proximity of the reference state to the quasi-chemical approximation is demonstrated in Figure 1D.

From this exercise, we learn the following: (1) QC appears to be in the neighborhood of the local optimal reference state. Indeed, it has been demonstrated¹³ that, in the situation of gapless threading of sequences through a diverse ensemble of conformations that preserve their native contacts, the mole-fraction product adequately approximates the probability of two amino acids to be in contact. This is consistent with Figure 1C, which demonstrates (via Eq. (15)) that QC optimizes the decoy term, thereby confirming that $p_{QC}^{exp}(c|ab) \approx p^D(c|ab)$. (2) The farther the reference states are from the quasi-chemical approximation, the lower the value of decoy term. Moreover, there are only 18 reference distributions, out of 400 randomly perturbed distributions, that produce a marginally higher score gap $J_q(C, S)$. (3) With reference to threading performance, we observe that the farther a probability distribution is from QC, the less effective it is in identifying the native state (Figure 1D).

3.2. Variations to the Quasi-Chemical Contact Reference State

The previous section pointed to the effectiveness of the quasi-chemical approximation in modelling the reference state. While there may be randomly perturbed reference states that can outperform QC, they do only by a small margin. More importantly, such reference states may prove impractical and undesirable because they don't arise from well-defined models or exact prescriptions. In this section, we confine our analysis to the space around probability distributions that arise from conceptually modelable states.

3.2.1. Variants of the quasi-chemical reference state—Apart from QC, a number of reference states have been advanced in the literature. Such models attempt to take into account relevant structural properties of natural proteins that QC does not. In particular, QC assumes that residues are not linked in chains of finite length, whose composition can differ significantly from the over-all composition of the “amino acid gas” (i.e., the composition of the universe of protein structures).³ More sophisticated reference state models take real-world characteristics of native protein conformations into account, in order to better estimate the “expected” probabilities of contact.¹⁰

There are a number of reference state models that attempt to consider the biases in amino acid composition within individual sequence chains of finite length.¹⁰ The first model we consider takes the expected probability of finding the pair ab in contact to be proportional to the number of times they exist together in the same sequence. The probability distribution can be derived from a set of sequences by the following formulation:

$$p_{QC1}^{\text{exp}}(c|ab) = \begin{cases} \frac{\sum_i 2n_i(a)n_i(b)}{\sum_i N_i(N_i-1)}, & a \neq b \\ \frac{\sum_i n_i(a)(n_i(a)-1)}{\sum_i N_i(N_i-1)}, & a=b \end{cases} \quad (22)$$

where N_i is the sequence length of sequence i . We shall refer to this reference state as QC1. Note that this expression reduces to QC once model-specific details are taken out. If amino acid composition $n_i(x)$ is assumed to be constant across all protein chains, the expression for $a \neq b$ in Eq. (22) becomes:

$$\frac{\sum_i 2n_i(a)n_i(b)}{\sum_i N_i(N_i-1)} = \frac{n_{\text{prot}} \cdot 2n(a)n(b)}{n_{\text{prot}} \cdot N_i(N_i-1)} \quad (23)$$

where n_{prot} is the number of chains in the data set. Upon applying the limit $N_i \rightarrow \infty$ (thus removing the detail that sequence lengths are finite), QC1 reduces to QC (Eq. 19) with $k = 2$. The same operation can be applied to the second expression ($a = b$), with the same result (with $k = 1$).

The second reference state (QC2) considers more specific structural properties of folded proteins. In this model, the contact probability is estimated as the mean of the expected probability of ab contact for each chain in the data set. This is calculated as follows:

$$p_{QC2}^{\text{exp}}(c|ab) = \frac{1}{\sum_i \sum_x \sum_{y \geq x} n_i^c(xy)} \sum_i f_i(ab) \sum_x \sum_{y \geq x} n_i^c(xy) \quad (24a)$$

where $n_i^c(ab)$ is the number of contacts between amino acids a and b in protein chain i , and

$$f_i(ab) = \begin{cases} \frac{2n_i(a)n_i(b)}{N_i(N_i-1)}, & a \neq b \\ \frac{n_i(a)(n_i(a)-1)}{N_i(N_i-1)}, & a=b \end{cases} \quad (24b)$$

Unlike QC1, QC2 recognizes that different folds occur in the data set, implying a variation in the number of contacts from chain to chain. The model, which has been referred to as partial-composition corrected reference state,¹⁰ proportionally partitions the number of contacts observed in *each* protein chain among the residue pairs within that chain, after which a weighted sum across all the chains in the data set is derived, to give the over-all expected probability. QC1, on the other hand, simply collates the proportion of expected pairings in the entire data set, without regard to fold detail. Mathematically, QC2 reduces to QC1 if the total number of contacts for each chain i is made proportional only to sequence

length. This is equivalent to setting $\sum_x \sum_{y \geq x} n_i^c(xy) = kN_i(N_i-1)$, at constant k , thus transforming Eq. (24a) into Eq. (22).

Results from the comprehensive all-against-all threading using QC and the two variant models QC1 and QC2 are summarized in Table I. Two sequence lengths ($L = 150, 200$) were used, along with two contact distances ($d_c = 9.5, 12.5\text{\AA}$), in order to survey a range of threading conditions. The difference in reference state models is clearly reflected by the decoy divergence term D . Incorporating progressively greater real-world detail into the model ought to bring the probability distribution $p^{\text{exp}}(c|ab)$ closer to what is actually observed in the data set $p^D(c, s)$. Although the inequality in Eq. (15) only deals only with the relationship between $p^D(c, s)$ and a given probability distribution, it should be possible to infer, from the results described here, the relationship among various probability distributions as a function of their distances from $p^D(c, s)$. In particular, the “closer” $p^{\text{exp}}(c|ab)$ is to $p^D(c, s)$, the higher the resulting D . Mathematically, $p^x(c_s, s_t)$ is “more proximate” to $p^D(c_s, s_t)$ than $p^y(c_s, s_t)$, when

$$\sum_{r,t} p^D(c_r, s_t) \log \frac{p^x(c_r, s_t)}{p^N(c_r, s_t)} > \sum_{r,t} p^D(c_r, s_t) \log \frac{p^y(c_r, s_t)}{p^N(c_r, s_t)} \quad (25)$$

Though divergence is not a distance in the strict sense, this relationship is useful in understanding the way reference state models act with respect to the amount of information incorporated in them. Any detail that can bring the expected probabilities of pairwise contact closer to what is actually seen in a typical threading exercise should increase the decoy divergence D .

Closer inspection of the data from all four threading sets in Table I, however, reveals that an increase in D does not necessarily mean a marked improvement in performance. The three models seem to perform similarly, with QC exhibiting slightly higher $\langle r \rangle$ than its two variants. The mutual information I is highest for QC, which more than offsets any drop in D to make its J maximal among the three models. These results suggest that the quasi-chemical approximation does at least as well as any of the more sophisticated models, and may even actually outperform them. These issues will be explored further in a later section.

3.2.2. Data-based reference state—The most accurate reference state model for a given data set can be computed directly from empirical pairing frequencies generated by the specific threading procedure. In this “data-based” model (DB),¹³ the expected probability of contact for any given pair is derived directly by aligning a series of sequences with a range of decoy conformations, and tallying every occurrence of the contact. In the context of a set of query protein sequences, the best estimate should be achieved by counting the total pairwise contact frequencies when *all* sequences are mounted onto *all* conformations. In effect, the empirical $p^D(c, s)$ is used to build the $p^{\text{exp}}(c|ab)$ of the score function, with the effect of achieving the equality demanded by Eq. (18). Because the quantities that make up this reference state are derived from direct counting of raw data, there is no mathematical expression (analogous to in Eqs. (22) and (24)) that can summarize it. The correlations among the scores resulting from the four models discussed thus far, summarized in Table II and illustrated in Figure 2, show that the differences among them are slight.

Results of comprehensive threading for this model are summarized in Table I. In accordance with Eq. (15), the decoy divergence D is highest for DB, continuing the trend established by QC1 and QC2. However, the decrease in mutual information I is more dramatic than the improvement in D , lowering the total divergence J , and resulting in a decreased average performance, as measured by $\langle r \rangle$.

3.2.3. Information-based comparison of reference states—We examine the relationships among DB, QC, and its two variants QC1 and QC2 more closely. If DB indeed embeds many characteristics of native-like chains in the model, then its probabilities $p_{DB}^{exp}(c|ab)$ should be closer to the true native probabilities $p^{obs}(c|ab)$ than $p_{QC}^{exp}(c|ab)$. This can be confirmed by a simple calculation. For each ab pair, relative distances among the three probabilities can be compared by:

$$\Delta = |p_{QC}^{exp}(c|ab) - p^{obs}(c|ab)| - |p_{DB}^{exp}(c|ab) - p^{obs}(c|ab)| \quad (26)$$

A positive Δ indicates that $p_{DB}^{exp}(c|ab)$ is closer to $p^{obs}(c|ab)$ than $p_{QC}^{exp}(c|ab)$, while a negative value indicates the opposite. Figure 3 shows that more than 86% of the unique amino acid pair probabilities that make up DB have positive Δ values, demonstrating that DB indeed exhibits more native-like character than QC. Likewise, comparisons between QC and QC1, and between QC and QC2, yield the expected ordering, namely that $QC < QC1 < QC2 < DB$ in terms of proximity to $p^{obs}(c|ab)$.

In constructing models that best embody the idea of “expected” probabilities, we seek ways to encode more native-like characteristics in the reference state. The limit of this exercise is the point where the reference state model approaches the observed contact probabilities, or $p_{QC}^{exp}(c|ab) = p^{obs}(c|ab)$. At this limit, Eq. (17) yields a value of zero for the three informatic quantities I , D , and J . Models that encode varying amounts of nativeness, including those examined here thus far, are informatically located between this extreme and QC. In order to examine the characteristics of such models, we built 100 evenly spaced reference state models from a weighted sum of $p_{QC}^{exp}(c|ab)$ and $p^{obs}(c|ab)$:

$$p_n^{exp}(c|ab) = \frac{1}{100} [n p_{QC}^{exp}(c|ab) + (100 - n) p^{obs}(c|ab)] \quad (27)$$

where $n = \{0, 1, 2, \dots, 100\}$, and subjected each to the same all-against-all threading. We note that this group of models is but a small subset of reference states that occur in this region. Random perturbations of any of the models, similar to the procedure in Section 3.1, reveals that the reference states generated by Eq. (27) are local optimal models at the particular level of “nativeness” (i.e., distance from $p^{obs}(c|ab)$) (results not shown). Therefore, consideration of the 100 models here should serve to evaluate locally optimal models in this region.

The “complete information” limit, described above, occurs at $n = 0$, while the QC is generated at $n = 100$. The informatic quantities that result from this set of models are plotted in Figure 4, spanning the range bounded by the state labelled “A” (at $n = 0$) to state “B” (at $n = 100$). In the right half of the figure, another 100 models were generated in a similar fashion, but this time forming a gradient from $p_{QC}^{exp}(c|ab)$ to the uniformly distributed reference state:

$$p_U^{exp}(c|ab) = \begin{cases} \frac{1}{200} & \text{for } a \neq b \\ \frac{1}{400} & \text{for } a = b \end{cases} \quad (28)$$

This reference state, marked as state “C” in the figure, assumes equal probabilities of finding any pair of amino acids. This is the extreme case of “ignorance”, in which even the most basic information, the uneven composition of the sequence universe, is not taken into

account. This is obviously not a practical nor acceptable model, and is included here only to serve as a limit.

The four reference states studied thus far are included in Figure 4, with the location of QC indicated by the dashed vertical line. First, we observe that the score gap $J_q(C, S)$ is maximal in the vicinity of QC, and drops as one proceeds from state B towards state A, passing through the three other reference state models along the way. The decoy term, whose negative value ($-D$) is tracked by the lower part of the plot, increases only slightly, its bottom remaining relatively flat over a wide range of reference states. This is consistent with the observation in the previous section that any benefit in the optimization of the decoy term is more than offset by the decrease in the native term (I), resulting in a depressed score gap $J_q(C, S)$.

The right side of the plot illustrates what happens when less and less prior knowledge is used to construct the reference state probabilities. Any attempt to increase “information” (as measured by the native term), by lowering the prior knowledge level, is offset by a proportional decrease of the decoy term, to produce a nearly constant $J_q(C, S)$ across the range of reference states. It is clear that these reference states are not useful in discrimination.

3.3. Contact Mole Fraction as Reference State

We examine a concrete example of a reference state that incorporates significantly more native characteristics (i.e., a model located in the left side of the information map in Figure 4). This reference state utilizes the quasi-chemical approximation not on amino acid composition (QC) but on the *contact* mole fraction. That is, the expected probability of contact between a and b is assumed to be proportional to the product of their individual contact mole fractions.

$$p_{\text{CQC}}^{\text{exp}}(c|ab) = k \chi_a^c \chi_b^c = k \left(\frac{\sum_i n_i^c(a)}{\sum_i N_i^c} \right) \left(\frac{\sum_i n_i^c(b)}{\sum_i N_i^c} \right) \quad (29)$$

where $n_i^c(a)$ is the number of contacts of a in protein chain i , $k=2$ if $a \neq b$ and 1 otherwise, N_i^c is the total number of contacts in protein i , and the summation covers all protein chains in the data set. This reference state, which we shall call here as CQC (as a reminder that this is the quasichemical approximation applied to contact mole fraction), is analogous to the GKS scale.^{13, 29}

While superficially similar, CQC and QC (Eq. (19)) differ significantly in the use of information. The latter uses the amino acid mole fraction χ_a , while the former uses the contact mole fraction χ_a^c . The difference arises from the fact that χ_a^c is dependent not only on the mole fraction of a , but also on its proclivity to be in contact. Thus, two amino acids a and b with the same χ can have significantly different χ^c if one tends to be in contact more than the other. A major consequence of employing this reference state model is the assignment of high $p_{\text{CQC}}^{\text{exp}}(c|ab)$ to amino acids that prefer to be in the protein interior, where pairwise contacts are more numerous than on the surface. This model assigns low $e_q(c|ab)$ for such hydrophobic pairs in native folds, compared to scores assigned by functions that use the other models discussed thus far. The converse is also true: polar amino acids, which are assigned low $p_{\text{CQC}}^{\text{exp}}(c|ab)$ by this model, will have higher $e_q(c|ab)$ for native conformations compared to the action of other score functions.

Operationally, using the CQC model has the effect of disregarding the influence of hydrophathy in the contact potential. Viewing this phenomenon in terms of information, QC-based potentials includes both the information on intrinsic contact propensities contained in CQC-based potentials as well as the information contained in the hydrophathy of individual amino acids. Thus, the CQC reference state can be said to hold significantly more native properties than QC, and therefore should be expected to occur in the left side of the information map in Figure 4.

The low correlation between CQC and the QC variants (Table II) and the plot comparing their $e_q(c|ab)$ values (Figure 2) both confirm that this model is fundamentally different. We can take the comparison further by quantifying the difference in information. If $e_q(c|ab)$ is a measure of the *specific* information content brought about by an ab contact, then the difference between the $e_q(c|ab)$ values given by CQC and QC should be a measure of the information in QC that is absent from CQC. The missing information involved with each of the 20 amino acids can be computed using a weighted mean of the difference:

$$h(a) = \frac{1}{\sum_x p^{\text{obs}}(c|ax)} \sum_x p^{\text{obs}}(c|ax) \left[e_q^{\text{QC}}(c|ax) - e_q^{\text{CQC}}(c|ax) \right] \quad (30)$$

Values for the $h(a)$ index can be found in Table III, along with three representative hydrophobicity/hydrophathy indices taken from the literature^{30–32} for comparison. The strong correlations among them, summarized in Table IV, confirm that the quantity $h(a)$ is, indeed, both a kind of data-derived hydrophobicity index, as well as a measure of the amount of information incorporated in score functions that use variants of QC but not those that use CQC.

Comprehensive threading results under the CQC model are summarized in Table I. Because hydrophathy is no longer incorporated in the resulting contact potential, the informatic quantities I , D , and J are significantly lower than those of the QC models. Indeed, these numbers show that CQC occurs in the left side of the information map in Figure 4. Consequently, its performance, as measured by $\langle r \rangle$, is diminished. These observations are expected. Score functions that use QC variants have been designed with the explicit purpose of summarizing all sequence-dependent pairwise contact information, including transfer energy or hydrophathy information. CQC, on the other hand, concerns itself only with actual pairwise propensities, without regard to location in the folded protein. This potential embodies only specific residue-residue interactions and excludes any effects arising from the collapse of the chain and the competitive interactions with respect to the aqueous solvent. Since the former set of score functions have included hydrophathy as a factor, they ought to perform significantly better in threading.

This does not mean that QC should always be chosen over CQC in constructing folding potentials. In building total potentials representing all the significant interactions which determine native conformation, surface/solvent effects can be included explicitly to account for the significant preference of hydrophobic residues to be buried in the protein interior. A separate term for this factor, in addition to a QC-type contact potential, may overcount burial effects in the total potential. Weighting parameters are used routinely to calibrate total potentials and limit such overcounting. But perhaps using CQC as the reference state for the contact potential term will ensure the independence of the contact and burial terms, minimizing the need for empirical weighting factors in the total potential expression. Further study is needed to uncover redundancies and untangle the elements which should be included in a total potential function.

3.4. Action of the Score Function on Individual Protein Chains

The information-theoretic dissection of the score-gap function above, as it relates to the *average* threading behavior across a range of test sequences and decoy conformations, can also guide us in the interpretation of the threading behavior of *individual* proteins.

The total divergence $J_q(C, S)$ can be recast as the score gap seen by each sequence s_i in the data set of n_{prot} protein chains:

$$J_q(C, S) = \frac{1}{n_{\text{prot}}} \sum_i J_q(C, s_i) \quad (31a)$$

where

$$J_q(C, s_i) = \sum_{\substack{\text{all pairs} \\ \text{in prot } i}} \frac{n_i^N(c|ab)}{n_i^N(c)} \log \frac{p^{\text{obs}}(c|ab)}{p^{\text{exp}}(c|ab)} + \sum_{\substack{\text{all pairs} \\ \text{in decoy} \\ \text{threading}}} \frac{n_i^D(c|ab)}{n_i^D(c)} \log \frac{p^{\text{obs}}(c|ab)}{p^{\text{exp}}(c|ab)} \quad (31b)$$

The same strategy used to study the optimization of $J_q(C, S)$ can be employed. In particular, the maximization of the native and decoy terms for *each* s_i , by equating database-specific and score function quantities (Eqs. (12) and (13)), suggests interesting directions. The native term of the score gap of protein i is maximized by the following equality:

$$\frac{n_i^N(c|ab)}{n_i^N(c)} = p^{\text{obs}}(c|ab) \quad (32)$$

while for the decoy term:

$$\frac{n_i^D(c|ab)}{n_i^D(c)} = p^{\text{exp}}(c|ab) \quad (33)$$

The quantities on the left hand side are properties of the query sequence s_i , while the quantities on the right hand side are components of the score function. The latter is derived typically from the statistics of a diverse pool of sequences and conformations, independent of the characteristics of the query sequence.

We begin by examining the decoy term equality first. The fraction on the left hand side is the proportion of contacts assigned to the ab pair in threading sequence i through the ensemble of decoy conformations. Using the quasi-chemical approximation, this fraction is proportional to the product of the mole fractions of a and b in protein i . On the other hand, also guided by the quasi-chemical approximation, the right hand side is proportional to the product of the mole fractions of a and b in the universe (U) of sequences. Thus, Eq. (33) can be approximated by the following:

$$\chi_i(a)\chi_i(b) = \chi_U(a)\chi_U(b) \quad (34)$$

for all ab residue pairs that exist in protein i . This system of 190 equations and 38 unknowns has one non-trivial solution:

$$\chi_i(x) = \chi_U(x) \quad (35)$$

for all amino acids x . Thus, *the decoy term is optimized if the amino acid composition of the data set used to construct the score function approximates that of the query protein i .* Enforcing this restraint may also provide the conditions to satisfy the first equality, Eq. (32). One way to allow for the possibility that the native distribution of contacts of a query sequence is similar to that found in the data set would be for both to have similar compositions. Conceptually, one can think of the native conformation as a coalescence of the most desirable (highest scoring) pairs in a competitive environment of a multitude of possible pairings. A different composition will alter the competitive environment, and therefore also the “winning” distribution of native contacts. In other words, the mutual information or “energy” of a particular amino acid pairing is a measure of its contact propensity in native conformations *relative* to the energies of the other contacting *and* non-contacting pairs. As a thought experiment, one may imagine observing exclusive binary contacts between an ab pair and a cd pair in a native conformation. Taking out the a residue from the sequence may affect the likelihood of the cd pairing, if alternative pairings bc or bd are more energetically stabilizing. Real proteins are certainly more complex than the binary contact situation just described, but one can hypothesize a similar high-order competition occurring in the collapse and rearrangement of residues in a folded chain.

We return to the comprehensive threading data to examine the validity of these propositions. First, we establish a definition of distance between the composition of a query sequence and that of the data set used to derive the score function. Again, the natural choice is the total divergence equation (Eq. (2)), which defines the distance between the compositions of sequence i and of the universe of sequences, X_i and X_U as follows:

$$J(X_i, X_U) = \sum_x \chi_i(x) \log \frac{\chi_i(x)}{\chi_U(x)} + \sum_x \chi_U(x) \log \frac{\chi_U(x)}{\chi_i(x)} \quad (36)$$

The value of the compositional divergence $J(X_i, X_U)$ for each 200-mer chain in the data set was plotted against its score gap $J_q(C, s_i)$ in Figure 5A. It should be expected that the bulk of the chains are found to be close in composition to the data set (i.e., low compositional divergence values), since the composition of the data set in fact arises from the aggregation of the data made up of the same proteins. However, there are a number of chains whose compositions diverge significantly from the data set, and these chains tend to exhibit lower score discrimination, as measured by $J_q(C, s_i)$. A plot of performance $\langle r \rangle$ and compositional divergence (Figure 5B) shows that the farther the compositional distribution of the test sequence is from the data set used to construct the score function, the lower the score gap $J_q(C, s_i)$, and the worse the score function does in discriminating the native conformation from the ensemble of decoys.

These measurements amplify an alternative approach to constructing potentials. There have been a number of studies that point to the influence of the data set on the performance of score functions in protein folding.^{33–36} If compositional divergence is partly responsible for the database dependence of the effectiveness of potential functions, a viable strategy would be to tailor potentials with respect to the composition of the query sequence. The idea is to select only chains in the structural database with low compositional divergence relative to the query sequence, in constructing the potential to be used to fold or thread that

sequence. The notion of “sequence-specific potentials” has been advanced by Skolnick and co-workers,¹⁰ who have weighted the contribution, to the potential, of each occurrence in the data set by the relative similarity of its local sequence to the query sequence. In this study, we expand the concept of query-specific potentials, by suggesting that composition, a global sequence characteristic, can also be used as a parameter to tailor potentials. Amino acid composition, of course, has been shown to contain information about over-all secondary structure content of proteins.^{37–39} In this work, we see the outline of an informatic connection between composition and long-range interactions.

We point out that we are pursuing issues relating to query-specific potentials in current work. Early results indicate that using only the most compositionally-similar part of the structural data set may, on average, be better than using the entire data set, despite the “representative” or nonredundant nature of the entire data set (Solis & Rackovsky, unpublished results).

4. CONCLUDING REMARKS

Using basic information-theoretic concepts, we set out to understand the role of the reference state in the action of folding potentials in fold recognition. We utilize information-theoretic quantities derived previously^{7, 8} to represent typical behavior of log-odds score functions in threading. We particularly study the total divergence, $J_q(C, S)$, which is the expected gap in scores between the native or correct conformation and the ensemble of decoy or incorrect conformations. It is useful to recall that total divergence is a sum of two terms (see Eq. (11)): the native term, which is also the mutual information brought about by the interaction being modeled by the potential, and the decoy term, which is the expected value of the mean of decoy scores. The score gap is seen to be a critical indicator of the performance of the potential function.

An explicit formulation of the discrimination score results in the establishment of a connection between the probability distributions which describe the reference state and the ensemble of decoy conformations. We find that the decoy term is maximized when these two distributions are identical. The farther the reference state probabilities are from the decoy ensemble probabilities, the lower the total divergence $J_q(C, S)$ and the higher the standard deviation of decoy scores $\sigma(D^*)$, resulting, on average, in diminished performance in fold recognition. These patterns become especially apparent when the distance between the distributions is significant. In practical terms, our results show that it is important that the reference state probabilities be modelled with the specific properties of the decoy ensemble in mind.

In the vicinity of the optimal reference state distribution, relationships are less well-defined. We examined the contact reference states derived from the quasi-chemical approximation (QC), the most commonly used model, and two variant models (QC1 and QC2). The basic quasi-chemical assumption takes the expected contact propensity to be proportional only to the mole fractions of the amino acids in the universe of natural sequences. The variants involve the inclusion of characteristics of protein chains: namely, that amino acid residues interact not in a “gas phase” but within chains of finite length, and whose amino acid distributions are variable. These variants, while increasing the decoy term of $J_q(C, S)$, decrease the native (mutual information) term. Consequently, we observe a slightly depressed $J_q(C, S)$ and performance, when measured by the expected rank $\langle r \rangle$ of native conformations.

A fourth reference state (DB) was derived directly using raw statistics, arising from the threading of a set of sequences through the ensemble of decoys. This kind of reference state realizes completely the equality between the probabilities of the decoy ensemble and the

reference state, and therefore produces the highest decoy term divergence. However, as in the case of the two quasi-chemical variants, the decrease in mutual information is larger, bringing about a lower $J_q(C, S)$ and higher $\langle r \rangle$. Among the models examined in this work, the quasi-chemical reference state QC proves to be as good a choice for reference state as any other. This is fortunate, as the QC model is also the easiest to formulate and implement.

We compared the informatic properties of the four reference states (QC, QC1, QC2, and DB). The difference among them lies in the amount of native-ness included explicitly in the model. In the attempt to approach the essential meaning of “expected” probabilities, more sophisticated models encode natural protein properties like finite length chains, uneven amino acid composition, and different conformations from chain to chain. It is conceptually appropriate to include such properties in the model. We find, however, that any amount of nativeness encoded in the reference state model does not significantly improve threading performance.

Another class of reference state models (CQC) uses the quasi-chemical approximation to partition the number of contacts according to the proclivity of the amino acids to be in contact in native folds. As a consequence, contacts made by amino acids that prefer to be in the protein interior receive lower scores than those given by QC-type models. This is because operationally, the prior expectation of contact for those amino acids have already been built into the score function. Conversely, contacts formed by amino acids found frequently on the protein surface are scored higher. Figure 2 illustrates the difference between CQC and QC. The preference to be on the protein surface or in the interior, of course, has been universally recognized as a consequence of the relative hydropathies of amino acid side chains. Thus, encoding these sequence-specific properties into the reference state has the effect of removing useful information about hydrophobic and solvent interactions from the contact potential. Results of all-against-all threading using CQC show that, indeed, the effectiveness of the score function is significantly depressed by the absence of hydropathy information in discrimination. We confirm the fundamental nature of the “missing” information further by building an amino acid index that quantifies the informatic difference between CQC and QC. This index, found in Table III, shows remarkable correlation with a number of hydrophobicity indices derived through other means. Though CQC-based contact potentials do not perform as well as QC-based ones, CQC may prove to be a useful component of a total potential if that potential also accounts for hydropathy and solvent accessibility explicitly.

The information map in Figure 4 is a useful guide to the range of reference states possible for contact potentials. The QC-based potentials are located at the region of greatest discrimination (i.e., highest J), which is consistent with their superior performance in threading. Reference states that contain varying amounts of “nativeness” are located in the left part of the map, among them the CQC reference state, which includes prior information about the relative hydropathy of amino acids (and therefore their location in the folded protein relative to the protein surface). Potentials which use these reference states see a depressed level of discrimination, owing to the increased prior knowledge contained in the base reference state. Conversely, located on the right side of the map are reference states that do not completely incorporate prior knowledge of the amino acid composition of natural proteins. It can be observed from this region of the information map that discrimination (as measured by J) cannot be artificially increased by any attempt to “increase” information (e.g., by removing prior information about amino acid composition from the reference state).

A promising avenue for the development of better potentials is suggested by our analysis of the action of the score gap function on *individual* protein sequences. If the same optimization is done to the two terms of this expression, an interesting prescription arises:

that the individual terms are optimized if the mole fractions of the decoy set and the reference state are equal. This result suggests that potentials have the capacity to perform better when the compositional properties of the data set used to derive the score function probabilities are similar to the properties of the sequence of interest. When one constructs the “expected” probabilities to form the reference state, one is really interested in the proportion of contacts that ought to be expected for the particular sequence being analyzed. Only similar sequences in the database will show the same “expected” contact behavior. Extracting parameters from similar sequences is also consistent with the real meaning of the other component of the score function: the “observed” contact probabilities. Sequences of near-identical amino acid compositions have available to them very similar possible permutations of contact (allowing for chain connectivity). Therefore, the set of contacts that finally “wins out” in the competition for the final conformation (i.e., the “observed” cases) should be a good indication of the relative “energies” of each contact pair. Our data supports these propositions remarkably well. We find that, the greater the divergence between the composition of the test sequence and the data set from which score parameters are derived, the lower the score gap, and the higher the mean rank $\langle r \rangle$ of the correct conformation. It may make sense, therefore, to use only sequences of similar composition to tailor the score function specifically for a test sequence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Library of Medicine of the National Institutes of Health, through grant LM06789.

References

1. 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. The calculation of low energy conformations from potentials of mean force. *J Mol Biol.* 1990; 216:167–80. [PubMed: 2121999]
2. Sippl MJ. Boltzmann’s principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des.* 1993; 7:473–501. [PubMed: 8229096]
3. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* 1995; 4:2107–17. [PubMed: 8535247]
4. Sippl MJ. Knowledge-based potentials for proteins. *Curr Opin Struct Biol.* 1995; 5:229–35. [PubMed: 7648326]
5. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. *Prot Struct Funct Bioinform.* 2000; 38:149–64.
6. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. *Prot Struct Funct Bioinform.* 2002; 48:463–86.
7. Solis AD, Rackovsky S. Improvement of statistical potentials and threading score functions using information maximization. *Prot Struct Funct Bioinform.* 2006; 62:892–908.
8. Solis AD, Rackovsky S. Information and discrimination in pairwise contact potentials. *Prot Struct Funct Bioinform.* 2008; 71:1071–1087.
9. 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; 256:623–44. [PubMed: 8604144]
10. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins.* 2000; 38:3–16. [PubMed: 10651034]

11. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A*. 2000; 97:2550–5. [PubMed: 10706611]
12. Miyazawa S, Jernigan RL. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*. 1985; 18:534–552.
13. 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; 6:676–88. [PubMed: 9070450]
14. Berrera M, Molinari H, Fogolari F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*. 2003; 4:8. [PubMed: 12689348]
15. Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. *Protein Sci*. 2005; 14:1741–52. [PubMed: 15987903]
16. Chelli R, Gervasio FL, Procacci P, Schettino V. Inter-residue and solvent-residue interactions in proteins: a statistical study on experimental structures. *Proteins*. 2004; 55:139–51. [PubMed: 14997548]
17. McConkey BJ, Sobolev V, Edelman M. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics*. 2002; 18:1365–73. [PubMed: 12376381]
18. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 2002; 11:2714–26. [PubMed: 12381853]
19. Liang S, Liu S, Zhang C, Zhou Y. A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. *Proteins*. 2007; 69:244–53. [PubMed: 17623864]
20. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*. 2004; 56:93–101. [PubMed: 15162489]
21. Cover, TM.; Thomas, JA. *Elements of Information Theory*. 2. Wiley; New York: 2006.
22. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*. 1990; 213:859–83. [PubMed: 2359125]
23. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol*. 1996; 6:195–209. [PubMed: 8728652]
24. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*. 2000; 10:139–45. [PubMed: 10753811]
25. Russ WP, Ranganathan R. Knowledge-based potential functions in protein design. *Curr Opin Struct Biol*. 2002; 12:447–52. [PubMed: 12163066]
26. Skolnick J. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol*. 2006; 16:166–71. [PubMed: 16524716]
27. Godzik A. Fold recognition methods. *Methods Biochem Anal*. 2003; 44:525–46. [PubMed: 12647403]
28. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol*. 1996; 264:1164–79. [PubMed: 9000638]
29. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol*. 1992; 227:227–38. [PubMed: 1522587]
30. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982; 157:105–32. [PubMed: 7108955]
31. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Prot Chem*. 1985; 4:23–55.
32. Casari G, Sippl MJ. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J Mol Biol*. 1992; 224:725–32. [PubMed: 1569551]
33. Dehouck Y, Gilis D, Rooman M. Database-derived potentials dependent on protein size for in silico folding and design. *Biophys J*. 2004; 87:171–81. [PubMed: 15240455]

34. Furuichi E, Koehl P. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins*. 1998; 31:139–49. [PubMed: 9593188]
35. Rykunov D, Fiser A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins*. 2007; 67:559–68. [PubMed: 17335003]
36. Zhang C, Liu S, Zhou H, Zhou Y. The dependence of all-atom statistical potentials on structural training database. *Biophys J*. 2004; 86:3349–58. [PubMed: 15189839]
37. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. *J Biochem*. 1986; 99:153–62. [PubMed: 3957893]
38. Eisenhaber F, Frommel C, Argos P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins*. 1996; 25:169–79. [PubMed: 8811733]
39. Eisenhaber F, Imperiale F, Argos P, Frommel C. Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. *Proteins*. 1996; 25:157–68. [PubMed: 8811732]

FIGURE 1.

Information-theoretic properties of 400 randomly generated reference states. These reference state distributions (δ) were generated by perturbation of the quasi-chemical reference state (QC). The distance between any reference state δ and QC can be computed using the total divergence expressed in Eq. (20). This distance is indicated in these figures by the quantity $J_q(p_{QC}^{exp}, p_{\delta}^{exp})$. Each of these reference states was incorporated in a contact potential, which was then used in an all-against-all threading exercise. (A) The mean gap score $J_q(C, S)$ and (B) $\sigma(D^*)$, plotted against the distance $J_q(p_{QC}^{exp}, p_{\delta}^{exp})$. (C) The negative of the scores given by decoys, or the negative of the decoy term, as expressed in Eq. (11), plotted against the distance $J_q(p_{QC}^{exp}, p_{\delta}^{exp})$. (D) The mean percentile rank $\langle r \rangle$ of the native conformation resulting from the use of the reference states, as a function of the distance of δ from the QC model.

FIGURE 2.

Scatter plots of the 190 score elements (one for each amino acid pair) of the contact potentials derived using the reference states examined in this work. (A) Comparison between the score elements given by the quasi-chemical reference state (QC) and the variant QC1, expressed in Eq. (22). (B) The score elements of QC1 and QC2, found in Eq. (24). (C) The score elements of QC and DB, the data-based reference state, as described in Section 3.2.2. (D) The score elements of QC and CQC, the contact mole fraction, as expressed in Eq. (29).

FIGURE 3.

Comparing the distance from $p^{\text{obs}}(c|ab)$, the true native probabilities, of two probability distributions given by the QC (quasi-chemical) and DB (data-based) reference states. The expression for the quantity Δ can be found in Eq. (26). If Δ is positive, then $p_{\text{DB}}^{\text{exp}}(c|ab)$ is closer to $p^{\text{obs}}(c|ab)$ than $p_{\text{QC}}^{\text{exp}}(c|ab)$, while a negative value indicates the opposite. In this plot, it can be observed that more than 86% of the unique amino acid pair probabilities that make up DB have positive Δ values, demonstrating that DB indeed exhibits more native-like character than QC.

FIGURE 4.

The information map, which explores the range of reference states, from the state encoding total knowledge of the contact probabilities (“A”) to the state encoding no knowledge (“C”). These reference states are generated using the procedure outlined in Section 3.2.3. The QC-type reference states (QC, QC1, QC2, and DB) are located in the vicinity of “B”, and are indicated by the black squares. The CQC reference state, located in the left side of the map, is indicated by black triangles. For each reference state, the three informatic quantities I , J , and $-D$ were calculated, and plotted in this figure. The x-axis is the I , the $-D$ are the empty circles (the bottom curve), and the J are the gray-filled diamonds (the top curve). Total knowledge is the state where the “expected” probabilities equal “observed” probabilities, while no knowledge is the state where the “expected” probabilities equal the uniform distribution (i.e., every amino acid pair has equal probability of contact).

FIGURE 5.

The action of the score function on individual protein chains. Each point in these two plots represents one protein chain. (A) The quantity $J(X_i, X_U)$ is the distance (measured by divergence) between the amino acid mole fraction composition of protein chain i and that of the entire data set. The quantity $J_q(C, s_i)$ is the gap between the score of the native conformation of chain i and the mean decoy score. While the bulk of the protein chains occur within a small area, there are some protein chains that diverge significantly from the data set, in terms of amino acid composition. These chains, with high $J(X_i, X_U)$, also have consistently low score gaps, as measured by $J_q(C, s_i)$. (B) Actual discrimination for each protein chain is measured directly by the mean threading rank $\langle r \rangle$. This plot confirms that the greater the divergence between the amino acid composition of the test protein from the aggregate amino acid composition of the data set, the poorer the discrimination in threading (i.e., higher native rank).

TABLE I
Information-Theoretic Quantities and Threading Performance of Contact Potentials with Varying Reference States

Parameters	QC	QC1	QC2	DB	QOC
I	6.42	5.93	5.89	5.47	1.14
$d_c = 9.5 \text{ \AA}$	-5.74	-5.89	-5.98	-6.09	-1.04
$L = 150$	12.16	11.82	11.87	11.56	2.18
$\langle r \rangle$	275	279	295	317	1869
$\langle Z \rangle$	4.02	4.07	4.02	4.05	2.73
I	6.09	5.58	5.50	5.00	1.09
$d_c = 9.5 \text{ \AA}$	-5.71	-5.88	-6.00	-6.17	-1.03
$L = 200$	11.80	11.46	11.50	11.17	2.12
$\langle r \rangle$	91	95	101	123	670
$\langle Z \rangle$	4.85	4.91	4.84	4.84	3.29
I	4.09	3.71	3.65	3.33	0.30
$d_c = 12.5 \text{ \AA}$	-3.60	-3.76	-3.84	-3.95	-0.34
$L = 150$	7.69	7.47	7.49	7.28	0.64
$\langle r \rangle$	273	279	302	328	3615
$\langle Z \rangle$	4.06	4.09	4.04	4.04	2.49
I	3.86	3.47	3.37	2.99	0.27
$d_c = 12.5 \text{ \AA}$	-3.59	-3.77	-3.88	-4.05	0.34
$L = 200$	7.45	7.24	7.25	7.04	-0.07
$\langle r \rangle$	102	106	118	144	1735
$\langle Z \rangle$	4.82	4.86	4.79	4.71	2.96

The information-theoretic quantities I , D , and J are all in units of cnats (1/100 of a nat). The mean native rank $\langle r \rangle$ is out of 58,034 150-mer decoy conformations and 37,727 200-mer decoy conformations. See text for a complete description of the five reference state models described in this table.

TABLE II

Correlation Between Energy Elements of Contact Potentials with Different Reference States

	QC	QC1	QC2	DB	CQC
QC	1.000				
QC1	0.983	1.000			
QC2	0.987	0.988	1.000		
DB	0.943	0.981	0.948	1.000	
CQC	0.376	0.317	0.297	0.325	1.000

TABLE III

Amino Acid Hydrophobicity/Hydropathy Scales

aa	h(a)	K-D	K	C-S
ALA	0.23	1.80	-0.27	0.20
CYS	0.85	2.50	-1.05	1.90
ASP	-0.85	-3.50	0.81	-1.40
GLU	-1.13	-3.50	1.17	-1.30
PHE	0.65	2.80	-1.43	1.00
GLY	-0.08	-0.40	-0.16	-0.10
HIS	-0.03	-3.20	0.28	0.40
ILE	0.67	4.50	-0.77	1.40
LYS	-1.01	-3.90	1.70	-1.60
LEU	0.45	3.80	-1.10	0.50
MET	0.34	1.90	-0.73	0.50
ASN	-0.58	-3.50	0.81	-0.50
PRO	-0.31	-1.60	-0.75	-1.00
GLN	-0.78	-3.50	1.10	-1.10
ARG	-0.50	-4.50	1.87	-0.70
SER	-0.29	-0.80	0.42	-0.70
THR	-0.09	-0.70	0.63	-0.40
VAL	0.69	4.20	-0.40	0.70
TRP	0.49	-0.90	-1.57	1.60
TYR	0.47	-1.30	-0.56	0.50

h(a) this work (Eq. 30)

K-D hydropathy index (Kyte-Doolittle, 1982)³⁰K hydrophobicity-related index (Kidera, et al., 1985)³¹C-S hydrophobicity scale from native protein structures (Casari-Sippl, 1992)³²

TABLE IV

Correlation Between Different Amino Acid Hydrophobicity/Hydrophathy Scales

	h(a)	K-D	K	C-S
h(a)	1.000			
KD	0.854	1.000		
K	-0.860	-0.768	1.000	
C-S	0.941	0.740	-0.817	1.000

h(a) this work (Eq. 30)

K-D hydrophathy index (Kyte-Doolittle, 1982)³⁰K hydrophobicity-related index (Kidera, et al., 1985)³¹C-S hydrophobicity scale from native protein structures (Casari-Sippl, 1992)³²