Inference of the solvation energy parameters of amino acids using maximum entropy approach

Trinh X. Hoang,[1,2] Flavio Seno,[3] Antonio Trovato,[4] Jayanth R. Banavar,[1] and Amos Maritan[3,a)]

[1]*Physics Department, Penn State University, 104 Davey Lab, University Park, Pennsylvania 16801, USA*
[2]*Instititute of Physics and Electronics, Vietnamese Academy of Science and Technology, 10 Dao Tan, Ba Dinh, Hanoi, Vietnam*
[3]*Dipartimento di Fisica "G. Galilei," Università di Padova and CNISM, Unità di Padova and INFN, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy*
[4]*CNISM, Unità di Padova and Dipartimento di Fisica "G. Galilei," Università di Padova, Via Marzolo 8, 35131 Padova, Italy*

(Received 11 April 2008; accepted 9 June 2008; published online 17 July 2008)

We present a novel technique, based on the principle of maximum entropy, for deriving the solvation energy parameters of amino acids from the knowledge of the solvent accessible areas in experimentally determined native state structures as well as high quality decoys of proteins. We present the results of detailed studies and analyze the correlations of the solvation energy parameters with the standard hydrophobic scale. We study the ability of the inferred parameters to discriminate between the native state structures of proteins and their decoy conformations. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2953691]

I. INTRODUCTION

The principle of maximum entropy is a powerful approach for statistical inference which is maximally noncommittal to missing information.[1–5] Our principal goal is to demonstrate the application of this technique to the vital problem of inferring the effective interactions between amino acids of proteins.[6–8] In the protein problem, the information that is readily available in the protein data bank[9] is the native state structures of a number of proteins determined from x-ray crystallography or NMR measurements.[10] An important, as yet unsolved, task is to develop procedures for predicting the native state structure of a new sequence of amino acids. Among the many routes for attempting to address this question is one requiring the development of a reliable scoring function that enables one to select the native state structure from among a lineup of candidate structures, often called decoys. Indeed, for several proteins in distinct classes, Levitt and co-workers[11–15] have developed a set of stringent decoys which can be used in the testing of the efficacy of scoring functions. The tests we carry out in this work are on such proteins and employ these decoys. For proteins without decoys, a powerful approach called threading[16] can be employed. This relies on the observation that the total number of distinct folds is limited[17–19] and thus pieces of known native state structures along with suitable insertions and deletions can be used as template structures for the selection process of the unknown native state structure.

There are several existing methods for deducing the scoring function from known protein structures and their sets of decoys. One of these, the quasichemical method,[20–23] is based on an assumption that the total number of contacts for each pair of amino acids observed in a set of many real protein native structures is equal to its expected value in an ensemble, in which all the contacts are found in quasichemical equilibrium with each other, i.e., the exchange of amino acids between the contacts is allowed. One may view this ensemble as a well-mixed gas of amino acids obtained by disassembling all the proteins together, and the probability of occurrence of the contacts for each pair of amino acids is determined by its Boltzmann weight. In addition, certain approximate schemes for capturing the chain connectivity may be implemented, e.g., via the coordination number for each type of amino acid and through the effective number of solvent molecules for each protein.[20,21] This method has the advantage of being easy to implement and no decoys are required. Nevertheless, it has an unpleasant feature of taking distinct systems, the different proteins, and artificially combining them into one supersystem, the gas of amino acids.[6,7,24]

A more stringent procedure was pioneered by Maiorov and Crippen,[25] who noted that one might "learn" the parameters of the scoring function by requiring that the native state structure has a better score than all the decoys for each of the proteins in a training set. This procedure entails the solution of a set of inequalities (one for each decoy) for the parameters in the scoring function.[26] A challenge for this method, apart from the difficulty of implementing it, is the identification of the correct parametrization of the scoring function. *A priori*, one does not know which kind of simple scoring function captures the essential features of the biology and is complete enough to ensure the existence of a solution. One often encounters an "unlearnable" situation in which there is no set of parameters which satisfies the requirement of the method.[27]

a)Electronic mail: maritan@pd.infn.it.

We will demonstrate that the principle of maximum entropy provides a sound basis for dealing with distinct proteins unlike the quasichemical method. The inputs to our calculation are constraints obtained from known data. One obtains a scoring function whose parameters are directly linked to these constraints. The maximum entropy method faithfully encodes these constraints which results in a decrease in the entropy of the system due to an increase in the information. In order to ensure that one is maximally noncommittal to missing information, one maximizes the entropy subject to the constraints because any entropy lower than this maximum permissible value would correspond to the incorporation of unwarranted additional information. Unlike the Maiorov–Crippen method, for which the problem may be unlearnable for some choices of the scoring function, the principle of maximum entropy yields a solution for the scoring function even when it is strictly unlearnable. The zero temperature limit of the maximum entropy method is related to the Maiorov–Crippen method.

The aim of our paper is to demonstrate and clarify the utility as well as the limitations of the maximum entropy method in specific examples. We will apply the maximum entropy method to infer the solvation energy parameters of amino acids from a knowledge of protein native structures and their decoy sets. There are two criteria we will use to test our results. The first is to compare the inferred solvation energy parameters to the hydrophobic scale from the literature—a significant correlation of the parameters with the hydrophobic scale would suggest that the inferred parameters are meaningful. The second criterion, which is more objective, is to check whether the inferred parameters can be used to discriminate between the native state and the decoy structures of proteins in the learning set as well as to predict the native state structures of proteins which are not in the learning set. We will assess the extent to which the maximum entropy method works well and provides a useful tool for deducing scoring functions for protein structure prediction.

II. MAXIMUM ENTROPY METHOD

We apply the maximum entropy (maxent) method to N_s protein sequences with given decoy sets. For each sequence we assume that the space of possible conformations is approximately captured by the corresponding set of decoy conformations together with the native state structure. Let Γ_{ia} $(0 \leqslant i \leqslant n_a)$ denote a decoy conformation for sequence $\hat{S}_a$ $(1 \leqslant a \leqslant N_s)$, where n_a is the total number of conformations in the decoy set of sequence $\hat{S}_a$. Γ_{0a} denotes the native state conformation.

Let us assume that we have a pool of N_s sequences, in which each sequence appears with equal probability, $q_a = 1/N_s$, whereas for a given sequence $\hat{S}_a$, the probability of adopting conformation Γ_{ia} is p_{ia}. The joint probability of selecting a sequence $\hat{S}_a$ in a conformation Γ_{ia}, from the pool, can be written as

$$P_q(\Gamma_{ia}, \hat{S}_a) = q_a p_{ia}. \tag{1}$$

The Shannon entropy is given by

$$\begin{aligned} \mathcal{S} &= -\sum_{a=1}^{N_s} \sum_{i=0}^{n_a} P_q(\Gamma_{ia}, \hat{S}_a) \ln P_q(\Gamma_{ia}, \hat{S}_a) \\ &= -\sum_{a=1}^{N_s} \sum_{i=0}^{n_a} q_a p_{ia} \ln p_{ia} - \sum_{a=1}^{N_s} q_a \ln q_a. \end{aligned} \tag{2}$$

The above entropy has a unique maximum with respect to the probability p_{ia} as it is a convex function of p_{ia} and its domain is compact.

Suppose that p_{ia}'s are unknown but they satisfy N constraints of the form

$$\langle \overline{C_\alpha} \rangle \equiv \sum_{a=1}^{N_s} \sum_{i=0}^{n_a} C_\alpha(\Gamma_{ia}) q_a p_{ia} = c_\alpha, \tag{3}$$

where $\alpha = 1, 2, \ldots, N$; $C_\alpha(\Gamma_{ia})$ are observable quantities associated with the ith decoy conformation of sequence $\hat{S}_a$; the averages c_α are hypothesized to be known a priori, $\overline{\cdots}$ denotes an average over conformations for each sequence and $\langle \cdots \rangle$ denotes an average over sequences. In addition, there are N_s constraints enforcing the normalization of the probabilities p_{ia} for each sequence,

$$\sum_{i=1}^{n_a} p_{ia} = 1. \tag{4}$$

The formulation of the maximum entropy method is to find the probabilities p_{ia}^* that maximize the entropy given by Eq. (2) while satisfying the constraints (3) and (4) exactly. This solution can be found through the use of N Lagrange multipliers, $\{\lambda_\alpha\} = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, to enforce the constraints (3) and N_s Lagrange multipliers, $\{\gamma_a\} = \{\gamma_1, \gamma_2, \ldots, \gamma_{N_s}\}$, to enforce the normalizations (4). One considers the following Lagrangian:

$$\begin{aligned} \Lambda(\{p_{ia}\}, \{\lambda_\alpha\}, \{\gamma_a\}) &= \mathcal{S} + \sum_{\alpha=1}^{N} \lambda_\alpha (\langle \overline{C_\alpha} \rangle - c_\alpha) \\ &\quad + \sum_{a=1}^{N_s} \gamma_a \left(\sum_{i=1}^{n_a} p_{ia} - 1 \right). \end{aligned} \tag{5}$$

On holding the Lagrange multiplier constant and maximizing the function Λ with respect to p_{ia}, and then by choosing γ_a to normalize p_{ia}, one obtains

$$p_{ia}(\{\lambda_\alpha\}) = \frac{1}{Z_a} \exp\left[\sum_{\alpha=1}^{N} \lambda_\alpha C_\alpha(\Gamma_{ia}) \right], \tag{6}$$

where

$$Z_a(\{\lambda_\alpha\}) = \sum_{i=0}^{n_a} \exp\left[\sum_{\alpha=1}^{N} \lambda_\alpha C_\alpha(\Gamma_{ia}) \right]. \tag{7}$$

With the solution given by Eqs. (6) and (7), the Lagrangian (5) transforms into

$$D(\{\lambda_\alpha\}) = \langle \ln Z \rangle - \sum_{\alpha=1}^{N} \lambda_\alpha c_\alpha, \tag{8}$$

where $\langle \ln Z \rangle = \Sigma_a q_a \ln Z_a$. It is straightforward to show that

TABLE I. Decoy sets used in the learning. The rate of successful prediction (fifth and sixth columns) presents the fraction of proteins whose native state has the lowest energy (rank=1) or is among five lowest energy conformations (rank ≤5) when using the Cornette hydrophobic scale as the solvation energy parameters in Eq. (12).

No.	Decoy set	Number of proteins	Average number of decoys per protein	Rate of successful prediction using hydrophobic scale[a]	
				rank=1	rank≤5
1	4state_reduced	7	665	2/7	5/7
2	fisa_casp3	5	1432	0/5	0/5
3	lattice_ssfit	8	2000	5/8	5/8
4	lmds	10	439	2/10	4/10
5	hg_structal	29	29	20/29	25/29
6	ig_structal	61	60	17/61	35/61
7	combined sets 1–4	26	1238	7/26	10/26
8	combined sets 1–6	116	317	44/116	70/116

[a]Cornette hydrophobic scale.

$$\langle \overline{C_\alpha} \rangle = \left\langle \frac{\partial \ln Z}{\partial \lambda_\alpha} \right\rangle = \frac{\partial \langle \ln Z \rangle}{\partial \lambda_\alpha}. \tag{9}$$

Thus, the constraints (3) are satisfied at the extremum of the function D. By minimizing D one obtains the solution $\lambda_\alpha = \lambda_\alpha^*$, which is unique given the convexity of the function itself. The probabilities p_{ia}^* that maximize the entropy $\mathcal{S}$ are obtained by applying the values λ_α^* into Eqs. (6) and (7). An inspection of the above equations shows that the λ's contain the Boltzmann factor $\beta = 1/k_B T$, so that their absolute values diverge in the $T \to 0$ limit. In this limit, it easy to see that the imposition of the constraints (3) is equivalent to satisfying the inequalities in the Maiorov–Crippen method.[25]

Let us assume that the energy of a protein in a given conformation can be parametrized as

$$E_{ia} = \sum_{\alpha=1}^{N} \epsilon_\alpha C_\alpha(\Gamma_{ia}), \tag{10}$$

and is measured in the units of $k_B T = 1$. For example, N could be 20, $\alpha = 1, 2, \ldots N$ could represent the types of amino acids, $C_\alpha(\Gamma_{ia})$ could be a measure of the solvent accessible area (SAA) of amino acid α when the sequence $\hat{S}_a$ is in conformation Γ_{ia}, and ϵ_α the solvation energy parameters for amino acid α. The probability p_{ia} in Eq. (6) becomes the Boltzmann weight of the conformation Γ_{ia} given that the Lagrange multipliers are now $-\epsilon_\alpha$. In order to apply the maxent method, one needs to know the averages c_α. The correct values of the latter, however, require a knowledge of the Boltzmann weights for each conformation, which are unknown. We make the plausible assumption of the thermodynamic dominance of the native state of proteins over the decoy conformations under physiological conditions.[28] In order to estimate the averages c_α, we postulate that all sequences have the same probability P for the native state and an equal probability $(1-P)/n_a$ for the decoys. Thus, c_α can be calculated as

$$c_\alpha = \sum_{a=1}^{N_s} q_a \left[P C_\alpha(\Gamma_{0a}) + \sum_{i=1}^{n_a} \frac{(1-P)}{n_a} C_\alpha(\Gamma_{ia}) \right]. \tag{11}$$

The above scheme has been successfully tested on a lattice HP model with the use of the numbers of contacts for the constraints. We note that although there are alternative schemes that one might consider, the maximum entropy method works well even in the case when the average numbers of contacts are not known with great accuracy.[29]

III. RESULTS AND DISCUSSION

A. Maxent with constraints employing SAAs

The SAAs of proteins were computed using the GETAREA software.[30] The observable quantities used for the constraints are the SAAs s_α in a given conformation for each type of amino acid α, where α can be one of the 20 amino acids. In the lowest order approximation, this corresponds to a total solvation energy of a protein in a given conformation Γ equal to

$$E(\Gamma) = \sum_{\alpha=1}^{20} s_\alpha(\Gamma) \epsilon_\alpha, \tag{12}$$

where ϵ_α is the amino acid dependent solvation energy parameters. It is straightforward to infer these energy parameters using the maxent method. We consider six decoy sets with known protein native structures provided by Levitt and coworkers[11–15] listed as the first six entries in Table I. We also consider two combined sets comprising all proteins in the first four sets and the first six sets in Table I. There are few proteins (1ctf, 1eh2, 1nkl, 1pgb, 2cro, 4pti) that appear in several sets. In the combined sets, all decoys corresponding to the same proteins are gathered together. We have computed the SAA for the native states of all proteins and all the decoys in the sets.

One can use the hydrophobic scales from the literature as the solvation energy parameters to test the energy of the native state against those of the decoys using Eq. (12). If the native state has lower energy than all the decoys, the result is

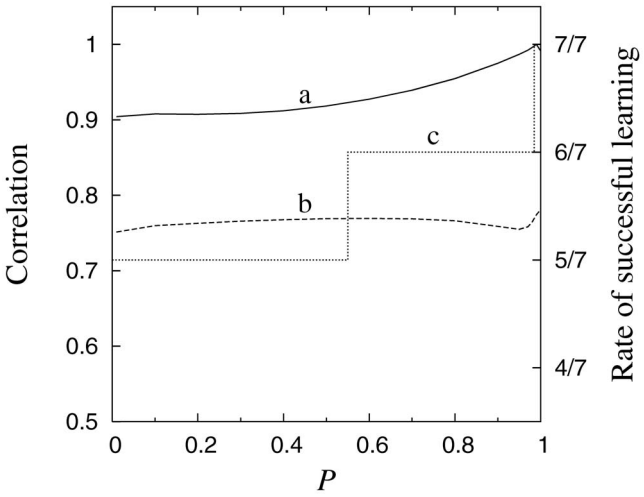

FIG. 1. Dependence of the maxent inference of 20 solvation energy parameters for the "4state_reduced" decoy set on the native probability P. (a) Correlation of the inferred parameters obtained at a given probability P with those obtained at $P=0.99$. (b) Correlation of the inferred parameters with the Cornette hydrophobic scale. (c) Rate of successful learning of the proteins in the decoy set.

considered successful. There are several hydrophobic scales in the literature. The best known include the Janin scale,[31] the Kyte–Doolittle scale,[32] the Hopp–Woods scale,[33] the Eisenberg scale,[34] the Rose scale,[35] the Engelman scale,[36] and the Cornette scale.[37] We find that the Cornette hydrophobic scale[37] is the best one in terms of distinguishing the native state structure from the decoy sets considered here. The fifth column of Table I shows the rate of successful prediction of the native state for proteins in each decoy set using the Cornette hydrophobic scale. We consider two criteria for our prediction: A strict one in which the native state energy is the lowest among the decoys (rank=1), and a less strict one in which the native state is found among five decoys of lowest energy (rank$\leqslant$5). The data show that in some cases, such as for the "lattice_ssfit" and the "hg_structal" sets, the prediction is surprisingly good. The Cornette scale was obtained by optimizing over 28 other published scales and is suitable for the prediction of α-helices in proteins. The other hydrophobic scales perform significantly worse (data not shown), and among these the Kyte–Doolittle scale is the second best.

We have found that the maxent procedure yields a robust solution with little dependence on the native state probability P. Figure 1 shows the P dependence for the inference of 20 solvation energy parameters using the "4state_reduced" decoy set. It is shown that the correlation between sets of the deduced parameters and between these and the hydrophobic scale is insensitive to P. The better scoring potential, in terms of the rate of successful discrimination between the native state structure and the decoys, is obtained with a higher P. We found that, when P is equal to or larger than 0.99, the native state structures of all proteins in the "4state_reduced" set are correctly "predicted," i.e., each of the native states has the lowest energy in comparison to the decoy energies. Going further, we will present our results with the value $P=0.99$. We will distinguish between two types of the rate of successful discrimination of the native states. When the inferred potentials are tested on the proteins in the training set, we call it the rate of successful learning, whereas when the potentials are tested on proteins other than in the training set, we denote our results in terms of the rate of successful prediction.

Table II summarizes the results for our maxent calculation with all the decoy sets. In most cases (except the "4state_reduced"), the parameters that correctly discriminate between true native state and the set of decoys are not completely learnable. Anyway, in contrast with the Maiorov and

TABLE II. Inference of the 20 solvation energy parameters using maxent with $P=0.99$. The rate of successful learning (third column) shows the number of proteins whose native state is ranked first (having the lowest energy) among the decoys in the learning set over the total number of proteins in the training set. The correlations of the inferred solvation energy parameters with the Cornette hydrophobic scale are given in the fourth column. The inferred energy parameters are then used to predict the native state of proteins which are present in the combined decoy sets 1–6 but *not present* in the training set. The rate of successful prediction (fifth and sixth columns) is defined as fraction of proteins successfully predicted using the inferred parameters. The prediction is considered as successful in two cases: When the native state has the lowest energy (rank=1) and when the native state is found among five conformations with the lowest energy (rank$\leqslant$5). The total rates of success are determined by applying the inferred energy parameters on all proteins in the "combined sets 1–6" and are given in the seventh and eighth columns.

No.	Decoy set used for learning	Rate of successful learning	Correlation of ϵ_α with hydrophobic scale	Rate of successful prediction[a]		Total rate of success[b]	
				rank=1	rank$\leqslant$5	rank=1	rank$\leqslant$5
1	4state_reduced	7/7	0.77	39/109	67/109	43/116	71/116
2	fisa_casp3	1/5	0.26	2/111	6/111	3/116	7/116
3	lattice_ssfit	6/8	0.74	20/108	45/108	25/116	51/116
4	lmds	9/10	0.22	3/106	14/106	9/116	21/116
5	hg_structal	27/29	0.57	16/87	33/87	43/116	62/116
6	ig_structal	45/61	0.52	3/55	8/55	48/116	68/116
7	combined sets 1–4	10/26	0.65	30/90	58/90	40/116	71/116
8	combined sets 1–6	67/116	0.72	0/0	0/0	67/116	93/116

[a]Tested on proteins in the "combined sets 1–6" and not present in the training set.
[b]Tested on proteins in the "combined sets 1–6."

Crippen approach, the maxent method allows always for an estimation of the parameters. In addition to the "4state_reduced" set, the rate of successful learning is quite high in other sets such as "lattice_ssfit," "lmds," and "hg_structal." The "fisa_casp3" is a challenging set because only one of the five proteins in the set is ranked first with the inferred energy parameters. For all the sets, the successful learning rate is always higher than the prediction rate using the hydrophobic scale. Table II shows the correlation between the energy parameters with the Cornette hydrophobic scale. This correlation varies between 0.22 (for the "lmds" case) and 0.77 (for the "4state_reduced" case). We have attempted to use the energy parameters deduced from one set to predict the native structure of proteins in the other sets. In order to have a consistent way for comparison, we will test the energy parameters on all the decoys in the "combined sets 1–6" for proteins that are not present in the training set to determine the rate of successful prediction. Our results are shown in the fifth and sixth columns of Table II. The highest and considerably significant rates of successful prediction are found for the three cases, "4state_reduced", "lattice_ssfit", and "combined sets 1–4." Interestingly, for these cases the correlations of the energy parameters with the hydrophbic scales are also the highest.

Note that the rates of successful prediction on using the Cornette hydrophobic scale are 44/116 and 70/116 for the "combined sets 1–6" for the two prediction criteria, rank $=1$ and rank $\leqslant 5$, respectively. The rates of successful prediction by using the inferred parameters generally are lower than those by using the hydrophobic scale. However, for the "4state_reduced" case, the performance of the inferred parameters (39/109 for rank$=1$ and 67/109 for rank $\leqslant 5$) is quite close to that by the hydrophobic scale.

Table II also shows the total rates of successful discrimination of the native state in both learning and prediction determined by applying the energy parameters to all proteins in the "combined sets 1–6" (seventh and eighth columns). For several cases, such as "4state_reduced", "hg_structal", and "ig_structal," the performance is comparable to that of the hydrophobic scale. The parameters deduced from the principle of maximum entropy perform better overall than the hydrophobic scale only when the "combined sets 1–6" are used as the training set (67/116 for rank$=1$ and 93/116 for rank $\leqslant 5$). The performance of the inferred parameters in the discrimination of the native state structure and the nature of the hydrophobic scale is shown in Fig. 2 for the above scoring function and other scoring functions to be considered hereon.

As a next order approximation, additional constraints were employed for not just the mean accessible area of the 20 amino acids but also on the values of the mean square accessible areas. This translates into a scoring function of the form

$$E(\Gamma) = \sum_{\alpha=1}^{20} [s_\alpha(\Gamma)\epsilon_\alpha^{(1)} + s_\alpha^2(\Gamma)\epsilon_\alpha^{(2)}], \tag{13}$$

where $\epsilon_\alpha^{(1)}$ and $\epsilon_\alpha^{(2)}$ are the solvation energy parameters of the first and second orders. We used maxent to deduce these 40

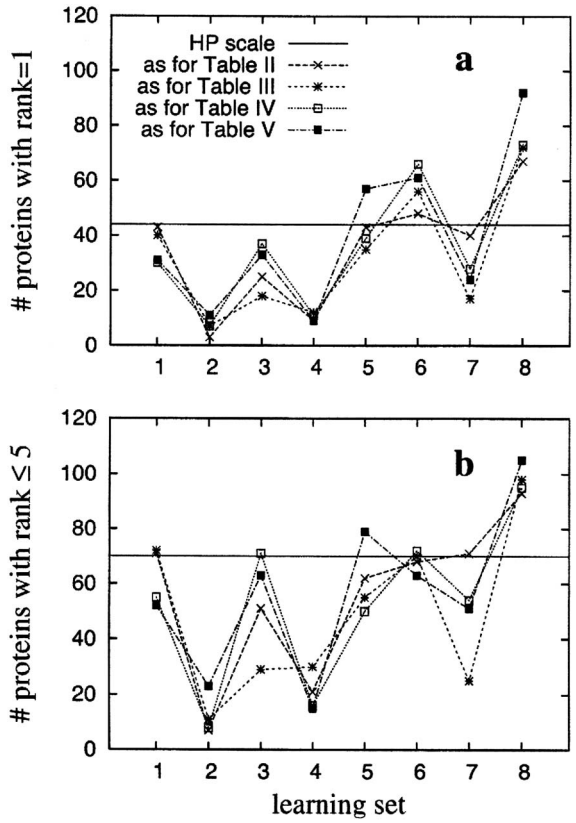

FIG. 2. The total number of proteins whose native state has the lowest energy (rank$=1$) (a) or is among the five lowest energy conformations (rank $\leqslant 5$) (b) using the inferred solvation energy parameters from eight different learning sets as shown in column 2 of Table I, and on employing different scoring functions as described in the text. The horizontal solid line indicates the performance of the hydrophobic scale as shown in Table I. The points connected by broken lines correspond to the data in the seventh (a) and eighth (b) columns of Table II–V, as indicated, and are obtained with different scoring functions given by Eqs. (12)–(15), respectively.

parameters, and Table III summarizes our results. It shows that on increasing the number of parameters the rate of successful learning improves significantly. In the first five of the total of eight different training sets, the rate of successful learning is 100%. However, the prediction rates decrease with increasing number of parameters. Overall, the combined rates of successful learning and prediction decrease in half of the cases and increase for the remaining cases. For the "combined set of 1–6" case, the total rates are 72/116 and 98/116 for rank$=1$ and rank $\leqslant 5$, respectively.

B. Maxent with additional constraints

Along with the constraints based on the SAAs, we also considered additional constraints employing other physical quantities. First, we considered constraints on both the average SAAs and the average number of contacts for each type of amino acid. This corresponds to a scoring function of the form

$$E(\Gamma) = \sum_{\alpha=1}^{20} [s_\alpha(\Gamma)\epsilon_\alpha^{(3)} + n_\alpha(\Gamma)\epsilon_\alpha^{(n)}], \tag{14}$$

where n_α is the total number of all contacts involving amino acid of type α in a given conformation Γ and $\epsilon_\alpha^{(n)}$ is a one-

TABLE III. Inference of the 40 solvation energy parameters $\epsilon_\alpha^{(1)}$ and $\epsilon_\alpha^{(2)}$ in Eq. (13) using the maxent procedure with $P=0.99$. The fourth column shows the correlations of the inferred solvation energy parameters $\epsilon_\alpha^{(1)}$ with the Cornette hydrophobic scale.

No.	Decoy set used for learning	Rate of successful learning	Correlation of $\epsilon_\alpha^{(1)}$ with hydrophobic scale	Rate of successful prediction[a]		Total rate of success[b]	
				rank=1	rank≤5	rank=1	rank≤5
1	4state_reduced	7/7	0.62	36/109	66/109	40/116	72/116
2	fisa_casp3	5/5	0.38	2/111	6/111	7/116	11/116
3	lattice_ssfit	8/8	0.78	11/108	22/108	18/116	29/116
4	lmds	10/10	0.32	5/106	23/106	12/116	30/116
5	hg_structal	29/29	0.49	6/87	26/87	35/116	55/116
6	ig_structal	53/61	0.05	3/55	10/55	56/116	70/116
7	combined sets 1–4	12/26	0.60	5/90	9/90	17/116	25/116
8	combined sets 1–6	72/116	0.49	0/0	0/0	72/116	98/116

[a]Tested on proteins in the "combined sets 1–6" and not present in the training set.
[b]Tested on proteins in the "combined sets 1–6."

body contact energy parameter for an amino acid of type α. A contact is defined to occur when two amino acids are in close proximity so that the distance between their C_α atoms is less than 7.5 Å. The 40 parameters ϵ_α and $\epsilon_\alpha^{(n)}$ are inferred using the maxent procedure. Table IV summarizes the results for $P=0.99$. It is shown that with the same number of constraints, the use of the average numbers of contacts is only marginally better than the use of the average squares of the solvation accessible areas. The former is comparable or better in five cases and is worse in the other three cases. On the other hand, the use of 40 parameters with the average numbers of contacts is comparable, in the overall rates of success, to the use of only 20 parameters with only the SAAs.

We also considered the maxent inference procedure for the solvation energy as well as local potentials related to bond and torsional angles. The latter ones accounts for the role of local steric interactions in proteins. Assume that a protein conformation is represented by N residues located at the positions of their C_α atoms. A bond angle of the ith residue in the chain is defined as the angle formed by three residues $(i-1,i,i+1)$. A torsional angle of residue i is defined as the angle between two planes defined by residues $(i-1,i,i+1)$ and $(i,i+1,i+2)$. A torsional angle is assigned a $+(-)$ sign according to the chirality of the local conformation. The constraints lead to a scoring function of the form

$$E(\Gamma)=\sum_{\alpha=1}^{20}[s_\alpha(\Gamma)\epsilon_\alpha^{(4)}+b_\alpha(\Gamma)\epsilon_\alpha^{(b)}+t_\alpha(\Gamma)\epsilon_\alpha^{(t)}], \tag{15}$$

where b_α and t_α are sums of the bond angles and torsional angles of all residues of type α in the sequence, respectively, and $\epsilon_\alpha^{(b)}$ and $\epsilon_\alpha^{(t)}$ are the energy parameters associated with them. Thus, one has 60 distinct parameters to be inferred from maximizing the entropy with 60 constraints. Table V summarizes our results. We note that the use of 60 parameters is much better than all previous tests only for the "hg_structal" and the "combined sets 1–6" case. For the latter, the total rates of successful discrimination of the native states are 92/116 and 105/116 for rank=1 and rank≤5, respectively. For other training sets, the performance of 60 parameters is comparable or only marginally better than 20 or 40 parameters. Table VI shows the values of inferred solvation energy parameters.

TABLE IV. Inference and results of testing of 40 parameters: 20 Solvation energy parameters $\epsilon_\alpha^{(3)}$ and 20 contact energy parameters $\epsilon_\alpha^{(n)}$ as defined in Eq. (14) were deduced using maxent with $P=0.99$.

No.	Decoy set used for learning	Rate of successful learning	Correlation of $\epsilon_\alpha^{(3)}$ with hydrophobic scale	Rate of successful prediction[a]		Total rate of success[b]	
				rank=1	rank≤5	rank=1	rank≤5
1	4state_reduced	7/7	0.66	25/109	50/109	30/116	55/116
2	fisa_casp3	5/5	0.21	2/111	3/111	7/116	8/116
3	lattice_ssfit	8/8	0.65	30/108	64/108	37/116	71/116
4	lmds	10/10	0.21	4/106	9/106	11/116	16/116
5	hg_structal	29/29	0.39	10/87	21/87	39/116	50/116
6	ig_structal	61/61	0.55	5/55	11/55	66/116	72/116
7	combined sets 1–4	9/26	0.71	19/90	40/90	28/116	54/116
8	combined sets 1–6	73/116	0.78	0/0	0/0	73/116	95/116

[a]Tested on proteins in the "combined sets 1–6" and not present in the training set.
[b]Tested on proteins in the "combined sets 1–6."

TABLE V. Inference of 60 parameters: 20 Solvation energy parameters $\epsilon_\alpha^{(4)}$ and 40 energy parameters related to the bond angle $\epsilon_\alpha^{(b)}$ and the torsional angles $\epsilon_\alpha^{(t)}$ as defined in Eq. (15) were deduced using maxent with $P=0.99$.

No.	Decoy set used for learning	Rate of successful learning	Correlation of $\epsilon_\alpha^{(4)}$ with hydrophobic scale	Rate of successful prediction[a]		Total rate of success[b]	
				rank=1	rank≤5	rank=1	rank≤5
1	4state_reduced	7/7	0.68	24/109	45/109	31/116	52/116
2	fisa_casp3	5/5	0.32	6/111	18/111	11/116	23/116
3	lattice_ssfit	8/8	0.81	26/108	56/108	33/116	63/116
4	lmds	10/10	0.06	2/106	8/106	9/116	15/116
5	hg_structal	29/29	0.72	28/87	50/87	57/116	79/116
6	ig_structal	61/61	0.14	0/55	2/55	61/116	63/116
7	combined sets 1–4	14/26	0.61	10/90	34/90	24/116	51/116
8	combined sets 1–6	92/116	0.70	0/0	0/0	92/116	105/116

[a]Tested on proteins in the "combined sets 1–6" and not present in the training set.
[b]Tested on proteins in the "combined sets 1–6."

IV. CONCLUSIONS

The maxent method can be used to derive the solvation energy parameters which is related to the hydrophobic scale by the use of SAAs. The parameters, not surprisingly, depend strongly on the decoy sets used in the learning. In some cases, the parameters show a significant transferability to proteins other than those in the learning set as well as an improvement with respect to the hydrophobic scale in terms of discriminating the native state structure from the set of decoys. We have shown that the maxent method provides a straightforward means of inferring the parameters of a scoring function in a variety of cases including constraints on the mean square of the solvation area of the amino acids, the number of contacts per amino acid, the local bond angles and torsion angles, etc. However, increasing the number of parameters does not necessarily improve the transferability of the potentials underscoring the fact that proteins in distinct classes do not necessarily share the same dominant energy contributions.

The maxent method is crucially dependent on three factors. The first is the parametrization of the scoring function to ensure the existence of a set of parameter values which is able to successfully discriminate between the native state structure and the decoys. The second is the choice of decoy conformations. They play a vital role in the determination of the partition function. The third is the crucial input of c_α in

TABLE VI. Values of the inferred solvation energy parameters for the 20 amino acids by the maxent method using the combined decoy sets 1–6 as the training set. The parameters shown are ϵ_α (second column) $\epsilon_\alpha^{(1)}$ (third column), $\epsilon_\alpha^{(3)}$ (fourth column), and $\epsilon_\alpha^{(4)}$ (fifth column) as defined in Eqs. (12)–(15), respectively. The parameters in each column are normalized in such a way that their rms value is the same as that of the Cornette hydrophobic scale (shown in the sixth column).

Amino acid	ϵ_α	$\epsilon_\alpha^{(1)}$	$\epsilon_\alpha^{(3)}$	$\epsilon_\alpha^{(4)}$	Cornette hydrophobic scale
ALA	1.10	1.73	1.31	2.40	0.20
ARG	1.02	−2.25	0.84	0.47	1.40
ASN	−1.16	−0.27	−0.62	−0.93	−0.50
ASP	−2.28	−6.83	−1.81	−1.50	−3.10
CYS	8.00	6.48	8.15	9.76	4.10
GLN	−3.14	0.18	−2.52	−1.18	−1.80
GLU	−0.66	1.89	−0.86	−0.51	−2.80
GLY	−2.86	0.57	−2.38	−2.36	0.00
HIS	−0.46	4.02	−0.33	−0.45	0.50
ILE	2.37	3.68	2.93	2.68	4.80
LEU	2.88	5.12	4.04	3.23	5.70
LYS	−0.57	−0.53	−0.33	−0.71	−3.10
MET	2.48	−0.04	3.65	2.87	4.20
PHE	3.76	3.82	4.10	3.52	4.40
PRO	−1.61	−0.01	−0.40	−0.53	−2.20
SER	−6.53	0.93	−5.39	−4.46	−0.50
THR	−2.33	2.10	−1.83	−1.68	−1.90
TRP	2.04	2.31	1.71	2.51	1.00
TYR	0.94	0.76	1.30	1.22	3.20
VAL	1.58	−1.26	2.52	1.70	4.70

Eq. (3). Because these averages are not easily deduced from existing data, here we chose to estimate c_α by using an *ad hoc* but plausible scheme of weighting the native state conformation with probability P (of the order of but slightly less than 1) and assigning equal but smaller weights to each of the decoy conformations. Careful attention to these three issues would lead to a more reliable scheme for using the maxent method in deducing scoring functions for proteins.

ACKNOWLEDGMENTS

The authors are indebted to Werner Braun and Numan Oezguen for making available the GETAREA source code. This work is supported by the National Program for Basic Research of Vietnam (Grant No. 404106). This article was funded in part by a grant from the Vietnam Education Foundation (VEF). The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of VEF. T.X.H. is a Vietnam Education Foundation visiting scholar.

[1] L. Boltzmann, *Lectures on Gas Theory* (Cambridge University Press, London, 1994).
[2] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[3] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[4] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[5] E. T. Jaynes, *Probability Theory* (Cambridge University Press, London, 2003).
[6] T. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996).
[7] F. Seno, A. Maritan, and J. R. Banavar, Proteins: Struct., Funct., Genet. **30**, 244 (1998).
[8] T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139 (2000).
[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[10] A. R. Fehrst, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
[11] B. Park and M. Levitt, J. Mol. Biol. **258**, 367 (1996).
[12] K. T. Simons, C. Kooperberg, E. S. Huang, and D. Baker, J. Mol. Biol. **268**, 209 (1997).
[13] Y. Xia, Y. E. S. Huang, M. Levitt, and R. Samudrala, J. Mol. Biol. **300**, 171 (2000).
[14] C. Keasar and M. Levitt, J. Mol. Biol. **329**, 159 (2003).
[15] R. Samudrala and M. Levitt, Protein Sci. **9**, 1399 (2000).
[16] D. T. Joens, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).
[17] C. Chothia, Nature (London) **357**, 543 (1992).
[18] M. Denton and C. Marshall, Nature (London) **410**, 417 (2001).
[19] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[20] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[21] S. Miyazawa and R. L. Jernigan, Proteins **34**, 49 (1999).
[22] M. J. Sippl, Curr. Opin. Struct. Biol. **5**, 229 (1995).
[23] R. Samudrala and J. Moult, J. Mol. Biol. **275**, 985 (1998).
[24] G. Salvi and P. De Los Rios, Phys. Rev. Lett. **91**, 258102 (2003).
[25] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).
[26] J. van Mourik, C. Clementi, A. Maritan, F. Seno, and J. R. Banavar, J. Chem. Phys. **100**, 10123 (1999).
[27] M. Vendruscolo and E. Domany, Proteins: Struct., Funct., Genet. **38**, 134 (2000).
[28] C. B. Anfinsen, Science **181**, 223 (1973).
[29] F. Seno, A. Trovato, J. R. Banavar, and A. Maritan, Phys. Rev. Lett. **100**, 078102 (2008).
[30] R. Fraczkiewicz and W. Braum, J. Comput. Chem. **19**, 319 (1998).
[31] J. Janin, Nature (London) **277**, 491 (1979).
[32] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[33] T. P. Hopp and K. R. Woods, Mol. Immunol. **20**, 483 (1983).
[34] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, J. Mol. Biol. **179**, 125 (1984).
[35] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, Science **229**, 834 (1985).
[36] D. M. Engelman, T. A. Steitz, and A. Goldman, Annu. Rev. Biophys. Biophys. Chem. **15**, 321 (1986).
[37] J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi, J. Mol. Biol. **195**, 659 (1987).