Maximum Entropy Approach for Deducing Amino Acid Interactions in Proteins

Flavio Seno,[1,2] Antonio Trovato,[1,2] Jayanth R. Banavar,[3] and Amos Maritan[1,2]

[1]*INFN and Dipartimento di Fisica, Università di Padova, Via Marzolo 8, I-35131 Padova, Italy*
[2]*CNISM, Unità di Padova, Via Marzolo 8, I-35131 Padova, Italy*
[3]*Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*

(Received 11 September 2007; published 20 February 2008)

We present a maximum entropy approach for inferring amino acid interactions in proteins subject to constraints pertaining to the mean numbers of various types of equilibrium contacts for a given sequence or a set of sequences. We have carried out several kinds of tests for a two-dimensional lattice model with just two types of amino acids with very promising results. We also show that the method works very well even when the mean numbers of contacts are not known and therefore can be applied to real proteins.

DOI: 10.1103/PhysRevLett.100.078102 PACS numbers: 87.15.Cc, 02.70.Rr

Globular proteins [1], the workhorse molecules of life, are linear chains of amino acids which fold rapidly and reproducibly into their native state conformations [2]. An important goal in the protein problem is the prediction of the native state structure given the amino acid sequence of a protein. A vast simplification arises because the number of distinct folds adopted by proteins is limited to just a few thousand [3]. This has led to a powerful method, called threading [4], for predicting the native state structure of a sequence—one mounts the sequence on candidate structures obtained as pieces of all known folds and determines the best fit structure through a scoring function [5–7], which provides a measure of the interactions between pairs of amino acids which are close by in a given conformation.

Our principal goal is to present and validate a knowledge-based method, using the principle of maximum entropy [8–10] for determining the scoring function subject to constraints pertaining to the mean numbers of various types of equilibrium contacts for a given sequence or a set of sequences. At zero temperature, this information is encoded in the native state structures of the sequences. In order to assess how well our method works and its ease of application, we will focus on extensive studies of the lattice HP model pioneered by Chan and Dill [11]. This model, while simple enough for exact enumerative studies, provides an unbiased framework for carrying out an extensive series of tests.

Entropy maximization has proved powerful in the derivation of equilibrium statistical mechanics [8], and in the analysis of complex equilibrium and nonequilibrium systems as neural networks [9] and global climate [10]. The underlying rationale is that each macroscopically observable state of a system corresponds to a number of microscopic states satisfying known macroscopic constraints. Because the number of ways of realizing a given macroscopic state can vary widely, the most likely state of the system as a whole is the one that corresponds to the largest number of microscopic states. As pointed out by Shannon [8], information and entropy are interlinked: the more information one has, the lower the entropy. The logic of our approach is to determine the scoring function subject to the entropy-reducing constraint that the available information on the mean number of certain contacts is faithfully encoded. Because the resulting scoring function is selected by the maximum entropy principle and assumes nothing about missing information, any system with lower entropy requires more information than is available from the given data.

Let $\mathcal{C}$ be the space of possible conformations and $P(\Gamma)$ the probability of a given sequence $\hat{S}$ adopting conformation Γ. The Shannon entropy is defined as

$$\mathcal{S}(P) \equiv -\sum_{\Gamma\in\mathcal{C}} P(\Gamma)\ln P(\Gamma). \tag{1}$$

Let the average values c_a of various observable quantities $C_a(\Gamma)$, $a = 0, 1, \ldots,$ be specified:

$$\overline{C_a} \equiv \sum_{\Gamma} C_a(\Gamma)P(\Gamma) = c_a. \tag{2}$$

This description includes the normalization condition, $\sum_\Gamma P(\Gamma) = 1$ on choosing $C_0(\Gamma) = 1$ and $c_0 = 1$. The maximum entropy principle states that the optimal choice for P is one that maximizes the entropy $\mathcal{S}$ while satisfying the constraints given by Eq. (2). This solution is unique. This is because the entropy $\mathcal{S}(P)$ is a convex function and its domain in P space is compact. Since the constraints are linear, from the theory of convex functions, $\mathcal{S}(P)$ has a single maximum. On using a set of N Lagrange multipliers ($\{\lambda_a\} = \lambda_1, \lambda_2, \ldots, \lambda_N$) to enforce the constraints, the solution takes the form

$$P(\Gamma) = Z(\{\lambda_a^*\})^{-1} e^{\sum_{a=1}^{N} C_a(\Gamma)\lambda_a^*}, \tag{3}$$

where

$$Z(\{\lambda_a\}) = \sum_{\Gamma} e^{\sum_{a=1}^{N} C_a(\Gamma)\lambda_a} \tag{4}$$

and the λ_a^*'s satisfy the equation

$$\overline{C_a} = \frac{\partial \ln Z}{\partial \lambda_a}\bigg|_{\lambda_a=\lambda_a^*} = c_a, \qquad a = 1, \ldots, N. \tag{5}$$

Numerically, an efficient way to obtain the solution of Eq. (5) is by finding the minimum of the quantity

$$D(\{\lambda_a\}) \equiv \ln Z(\{\lambda_a\}) - \sum_{a=1}^{N} c_a \lambda_a, \tag{6}$$

which is unique since $\partial^2 D/\partial\lambda_a\partial\lambda_b = \overline{(C_a - c_a)(C_b - c_b)}$ is positive definite.

The analysis can be extended to the quenched case of the simultaneous consideration of many sequences (we denote a sequence by $\hat{S}$). In this case the probability $P(\Gamma|\hat{S})$, the observables $C_a(\Gamma|\hat{S})$, and their average values over conformations $\overline{C_a(\hat{S})} = \sum_{\Gamma\in\mathcal{C}} C_a(\Gamma|\hat{S}) P(\Gamma|\hat{S})$ depend on another statistical variable, $\hat{S}$. The constraint is then imposed through the *quenched* average

$$\langle \overline{C_a} \rangle \equiv \sum_{\hat{S}} \overline{C_a(\hat{S})}\, q(\hat{S}) = c_a, \tag{7}$$

where $q(\hat{S})$ represents the probability of occurrence of the variable $\hat{S}$. This case can be mapped into the standard one by introducing the joint probability distribution

$$P_q(\Gamma, \hat{S}) \equiv P(\Gamma|\hat{S}) q(\hat{S}) \tag{8}$$

where $q(\hat{S})$ is given *a priori* and the entropy is

$$\mathcal{S}(P_q) = -\sum_{\Gamma,\hat{S}} P_q(\Gamma, \hat{S}) \ln P_q(\Gamma, \hat{S}) \equiv \langle \mathcal{S}(P) \rangle + \mathcal{S}(q). \tag{9}$$

Its maximum is given by the straightforward generalization of Eqs. (3) and (4), where the λ_a^*s satisfy the equation

$$\langle \overline{C_a} \rangle = \left\langle \frac{\partial \ln Z}{\partial \lambda_a} \right\rangle \bigg|_{\lambda_a=\lambda_a^*} = \frac{\partial \langle \ln Z \rangle}{\partial \lambda_a}\bigg|_{\lambda_a=\lambda_a^*} = c_a, \tag{10}$$

where

$$\langle \ln Z(\{\lambda_a\}) \rangle = \sum_{\hat{S}} q(\hat{S}) \ln Z(\{\lambda_a\}|\hat{S}). \tag{11}$$

This is (apart from the $-\kappa_B T$ term) the quenched free energy used in the standard approach of disordered statistical mechanics systems (see, e.g., [12]). Again the λ_a^*'s correspond to the unique minimum of the function

$$D_q(\{\lambda_a\}) \equiv \langle \ln Z(\{\lambda_a\}) \rangle - \sum_{a=1}^{N} c_a \lambda_a. \tag{12}$$

We now proceed to illustrate the method and carry out several tests of its efficiency within the framework of the HP lattice model in two dimensions [11]. It has been recognized that in proteins a simply binary pattern of hydrophobic and hydrophilic residues along the chain encode structure at the coarse grained level [13]. Thus the simplest model of proteins consists of sequences made up of just two kinds of amino acids (H and P representing hydrophobic and polar residues) configured as self-avoiding chains on a lattice and described by a contact Hamiltonian [11]. Such models are known to adequately describe proteins at the coarse-grained level with the advantage that the native states can be determined exactly. Furthermore, they provided a controlled laboratory for theoretical investigations and rigorous testing of concepts and ideas for future use in studies on real proteins [7,11,14]. We work with short chains constrained to lie on a square lattice. Two amino acids interact when they sit next to each other and are not contiguous along the chain, yielding a scoring energy function

$$\mathcal{H}_{nn}(\Gamma, \hat{S}) = -\sum_{a=1}^{3} \epsilon_a C_a(\Gamma|\hat{S}), \tag{13}$$

where the index $a = 1, 2, 3$ labels HH, HP, and PP type of interactions, respectively, and $C_a(\Gamma|\hat{S})$ counts the number of type a interactions when the sequence $\hat{S}$ is in the conformation Γ. In order to assess the effect of a redundant, incomplete, or wrong parametrization, we have considered a second scoring function to include next nearest neighbor interactions

$$\mathcal{H}_{nnn}(\Gamma, \hat{S}) = -\sum_{a=1}^{6} \epsilon_a C_a(\Gamma|\hat{S}), \tag{14}$$

where the index $a = 4, 5, 6$ labels next nearest neighbor interactions of type HH, HP, and PP, respectively. The hydrophobic nature of a protein is encapsulated typically by using a scoring function in which HH contacts are favored over HP and PP contacts. In our studies, we considered different choices of the interactions energies. Most of our studies were carried out with a chain of length $L = 16$ for which there is an ensemble $\mathcal{C}$ of 802 075 distinct (compact and noncompact) conformations Γ unrelated by simple symmetry transformations. We also considered chains of length $L = 12$ in additional tests. The Lagrange multipliers which minimize the functions D and D_q in Eqs. (6) and (11) are estimates of the interaction energies $\frac{\epsilon_a}{k_B T}$ in Eqs. (13) and (14).

We have carried out two classes of tests. The first entails (i) choosing the interaction energies, (ii) determining the Boltzmann probability distribution

$$P(\Gamma|\hat{S}) = \frac{\exp[\frac{-\mathcal{H}(\Gamma,\hat{S})}{k_B T}]}{\sum_{\Gamma'\in\mathcal{C}} \exp[\frac{-\mathcal{H}(\Gamma',\hat{S})}{k_B T}]} \tag{15}$$

and then the average number of different types of contacts in equilibrium at a given temperature through exact enumeration [using Eq. (2) for a single sequence and Eq. (7) for the quenched case, with $q(\hat{S}) = 1/N_s$ for N_s sequences], (iii) using the c_a as constraints in the maximum entropy approach to estimate the interaction energies and assess the quality of agreement of these estimates with the actual values (the scoring function $\mathcal{H}$ could be either $\mathcal{H}_{nn}$

TABLE I. Reconstruction of interaction energies with the maximum entropy procedure. These results were obtained for both a single, randomly selected sequence, and for ten different sequences in the quenched case. In all cases, the chain had a length of 16.

k_BT	ϵ_1	ϵ_2	ϵ_3	$k_BT\lambda_1^*$	$k_BT\lambda_2^*$	$k_BT\lambda_3^*$
1.0	1	0	0	1.0000	0.0000	0.0000
2.0	1	0	0	1.0000	0.0000	0.0000
0.5	1	0	0	1.0000	0.0000	0.0000
1.0	1	0.5	−0.2	1.0000	0.5000	−0.2000
2.0	1	0.5	−0.2	1.0000	0.5000	−0.2000
0.5	1	0.5	−0.2	1.0000	0.5000	−0.2000

or $\mathcal{H}_{nnn}$). We have carried out extensive studies and in *all cases* (single sequence, multiple sequences, either of the same or of different lengths, contact potentials excluding or including next nearest neighbors, calculations at different temperatures) the results are excellent with the deduced scoring function being virtually exact, as summarized in Tables I, II, and III.

The equilibrium stability of various contacts of a specific protein sequence or set of sequences could be in principle deduced by means of NMR experiments, but this would require the computation of the partition function (4) many different times within the entropy maximization procedure. More simply, one can test or deduce scoring function by ranking the native state against competitive conformations called decoys. In this spirit, we carried out a second class of tests of the HP model employing the knowledge of the ground state conformation for a given sequence as well information pertaining to the native state conformations for other sequences to be used as decoys. In order to obtain a measure of the mean numbers of different contacts, we postulated that a given sequence has probability P_0 to be in its native state and a probability $(1 - P_0)/(M - 1)$ (where M is the number of known ground state conformations in the data bank) of being in one of the other structures. We note that there are simple alternative schemes that one might consider—the key point is that the maximum entropy method works well even if the average numbers of contacts are not known with great accuracy.

We considered the ensemble $\mathcal{GS}$ of the nondegenerate conformations which are unique ground states for the HP model with nearest neighbor interactions and interaction strength $\epsilon_{\rm HH} = 5$, $\epsilon_{\rm HP} = 3$, $\epsilon_{\rm PP} = 1$. The number of such conformations is $M = 1131$ and the partition function Z was computed on the ensemble $\mathcal{GS}$, while implementing the maximum entropy procedure. We then used the solution λ_1^*, λ_2^*, λ_3^* and assessed how many of the 17 021 sequences with a unique ground state have their native state conformation as their ground state among all the 802 075 conformations (not just the ensemble $\mathcal{GS}$).

TABLE II. Reconstruction of interaction energies with the maximum entropy procedure in the quenched case with multiple sequences having different lengths. These results were obtained by averaging over 6 sequences of length $L = 16$ and 4 of length $L = 12$.

k_BT	ϵ_1	ϵ_2	ϵ_3	$k_BT\lambda_1^*$	$k_BT\lambda_2^*$	$k_BT\lambda_3^*$
1.0	1	0.5	−0.2	1.0000	0.5000	−0.2000
1.0	1.5	0.7	−0.2	1.5000	0.7000	−0.2000
1.0	1.66	0.41	0.14	1.6600	0.4100	0.1400

TABLE III. Reconstruction of interaction energies with the maximum entropy procedure in the presence of next nearest neighbor interactions for a single, randomly selected, sequence. In all rows, $\epsilon_1 = 1$, $\epsilon_2 = 0$, $\epsilon_3 = 0$, with a chain of length 16. In all rows, but the second one, $\epsilon_4 = 0.5$, $\epsilon_5 = 0.2$, $\epsilon_6 = -0.1$. In the first row, the full knowledge of the scoring function is used and all six λ^*'s are determined. In the second row, the next nearest neighbor interaction strengths are set equal to zero and all six λ^*'s are determined. In the third and fourth row, only the three λ^*'s corresponding to nearest neighbor interactions are reconstructed. Interestingly, when one uses a subset of constraints, the results are temperature dependent. This property can be used to assess whether our parametrization of the scoring function is complete or not. The $\{\lambda_\alpha^*\}$ are different from the original ones, as expected. Yet, the sequence with the reconstructed energies of interaction yield exactly the same ground state as the original scoring function with all 6 energy parameters.

k_BT	$k_BT\lambda_1^*$	$k_BT\lambda_2^*$	$k_BT\lambda_3^*$	$k_BT\lambda_4^*$	$k_BT\lambda_5^*$	$k_BT\lambda_6^*$
1	1.000	0.0000	0.0000	0.5000	0.2000	−0.1000
1	1.000	0.0000	0.0000	0.0000	0.0000	0.0000
1	1.8147	0.4642	0.1180			
2	1.7146	0.4808	0.1596			

Figure 1 reports the rate of success as a function of the probability P_0 of occupancy of the ground state for different cases. When single sequences were considered, the rate

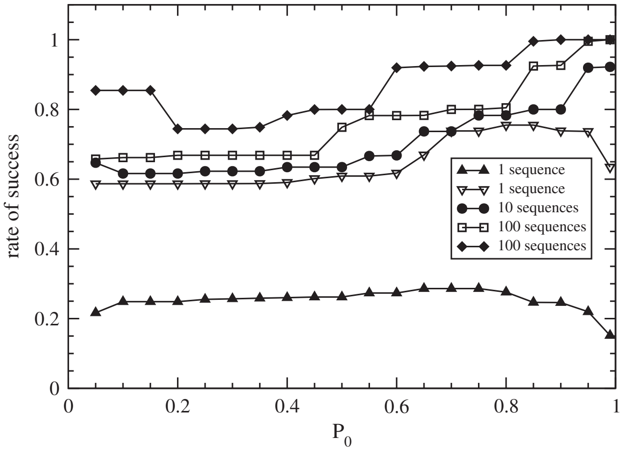

FIG. 1. Fraction of the 17 021 sequences, with a unique ground state ($\epsilon_1 = 5$, $\epsilon_2 = 3$, $\epsilon_1 = 1$) in Eq. (13), for which the maximum entropy method correctly predicts the true ground state as a function of the probability P_0. The different curves correspond to different sequence sets considered.

of success was strongly dependent on the chosen sequence. For a given sequence, the method worked with varying degree of success for an optimal value of P_0 for which the actual Boltzmann distribution obtained with the Lagrange multipliers [Eq. (3)] approximated our assumption of a probability P_0 of being in the ground state and an equal probability of all remaining conformations.

On averaging multiple sequences together (quenched case), one would expect an overall smoothing to occur so that, at low temperatures (high P_0), our assumption might well be satisfied approximately. This is indeed the case for the quenched case of 10 sequences and the situation is much improved when 100 sequences are considered. For two distinct sets of sequences, we obtained 100% success, for values of P_0 greater than 0.95, in predicting the native state structure of all the sequences from the deduced interaction energies.

In summary, we have presented a novel method for determining the all-important scoring function for the interaction between amino acids of a protein. The method entails the use of the principle of maximum entropy which provides an unbiased method for inferring the scoring function while incorporating all available information, thus entirely avoiding unwarranted assumptions which may be present in other approaches [15]. Detailed studies on a lattice HP model have yielded very promising results. Note that there are other methods which work well in the context of lattice models [6,7,16]. However, in such approaches, the energy parameters are typically obtained by ensuring that the native state conformation has a lower energy than the alternative conformations. In general, one encounters situations in which the problem is unlearnable [17] and no solution exists. In contrast, in our approach, one is guaranteed to find the optimal solution which satisfies physical constraints derived from experimental observations in terms of average values. This is quite important in light of one's increasing ability to perform single molecule experiments. Furthermore, the ease with which one can incorporate other information such as the solvent accessible area and/or local conformational biases make the maximum entropy inference (maxent) approach a very promising tool for study of the daunting protein problem. We have employed the procedure employed in the second class of tests to investigate a single protein sequence (PDB code 1CTF) together with 498 decoys from the local minima decoy set (lmds) [18]. By defining the scoring function in terms of solvent accessible areas, 20 surface tension parameters are recovered using the maxent method which correctly selects the native state conformation from among all the decoys [19].

This work was supported by PRIN (No. 2005027330).

[1] A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman and Company, New York, 1999).

[2] C. B. Anfinsen, Science **181**, 223 (1973).

[3] C. Chothia, Nature (London) **357**, 543 (1992); M. Denton and C. Marshall, Nature (London) **410**, 417 (2001); T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).

[4] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).

[5] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985); M. J. Sippl, Curr. Opin. Struct. Biol. **5**, 229 (1995).

[6] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).

[7] F. Seno, A. Maritan, and J. R. Banavar, Proteins: Struct. Funct. Genet. **30**, 244 (1998); J. van Mourik, C. Clementi, A. Maritan, F. Seno, and J. R. Banavar, J. Chem. Phys. **110**, 10123 (1999).

[8] L. Boltzmann, *Lectures on Gas Theory* (Cambridge University Press, London, U.K., 1964); C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); E. T. Jaynes, Phys. Rev. **106**, 620 (1957); E. T. Jaynes, Phys. Rev. **108**, 171 (1957); E. T. Jaynes, *Probability Theory* (Cambridge University Press, London, U.K., 2003).

[9] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).

[10] R. C. Dewar, J. Phys. A **36**, 631 (2003); R. C. Dewar, J. Phys. A **38**, L371 (2005).

[11] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989); K. A. Dill, S. Bromberg, S. Yue, K. Fiebig, K. M. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).

[12] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glasses Theory and Beyond* (World Scientific, Singapore, 1987).

[13] Y. Kuroda, T. Nakai, and T. Ohkubo, J. Mol. Biol. **236**, 862 (1994); T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, and D. C. Richardson, Proc. Natl. Acad. Sci. U.S.A. **91**, 8747 (1994); S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).

[14] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995); J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996); F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996); F. Seno, C. Micheletti, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).

[15] P. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996); G. Tiana, D. Colombo, D. Provasi, and R. A. Broglia, J. Phys. Condens. Matter **16**, 2551 (2004).

[16] G. Salvi and P. De Los Rios, Phys. Rev. Lett. **91**, 258102 (2003).

[17] M. Vendruscolo and E. Domany, Proteins: Struct. Funct. Genet. **38**, 134 (2000).

[18] R. Samudrala and M. Levitt, Protein Sci. **9**, 1399 (2000).

[19] T. X. Hoang, F. Seno, A. Trovato, J. R. Banavar, and A. Maritan (to be published).