Effective Interactions Cannot Replace Solvent Effects in a Lattice Model of Proteins

G. Salvi and P. De Los Rios

Laboratory of Statistical Biophysics, ITP-FSB, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

(Received 8 April 2003; published 16 December 2003)

Protein folding and protein design are among the most challenging problems of the past ten years in biophysics and molecular biology. For a given protein, it is possible to extract, from existing protein databases, a set of specific (i.e., belonging to the investigated protein) effective amino-acid (AA) interactions able to stabilize the native state. On the other hand, attempts to find global effective AA interactions, which would be able to stabilize all proteins at once, failed. Using a simple lattice model where the solvent degrees of freedom are (semi)explicitly taken into account, we show that the absence of global effective AA interactions is due to the solvent and that on this lattice model the solvent effects cannot be reproduced by amino-acid effective interactions.

DOI: 10.1103/PhysRevLett.91.258102 PACS numbers: 87.15.Aa, 87.14.Ee, 87.15.Cc

Protein folding has become one of the most challenging problems of the past ten years, involving biology, chemistry, medicine, and physics. It is believed that the 3D structure of a protein is completely encoded in its amino-acid (AA) sequence: this is the fundamental dogma in protein science [1], and protein folding research aims at discovering the right key able to decode the AA sequences. Not less important, for example, for drug engineering, is protein design: the prediction of the residue sequences $\{S\}$ that admit as their native state a predefined 3D conformation Γ. In principle, these problems could be tackled using all-atoms molecular dynamics, but this method has been successful only for small proteins [2]. Therefore, people have introduced models that reduce the complexity of the system to a manageable level. One usual strategy is to determine effective AA/AA contact interactions for coarse-grained models of proteins. These contact interactions can be encoded in 20×20 matrices (there are 20 naturally occurring residues) [3], and the native state corresponds to the minimum of a given cost function depending on them [4]. To determine these interactions several techniques have been developed, such as the PERCEPTRON algorithm [5], where a simple neural network tries to learn contact interactions from a given training set of sequences and structures (as, for example, from the Protein Data Bank) [6]. However, even if this technique has turned out to be one of the most promising methods, the computed effective interactions are able to stabilize only small sets of proteins [7]: it is not possible to determine global effective interactions that stabilize all proteins at once. Actually, in the worst case, a simple pairwise energy function may not stabilize even a single protein such as crambin, as shown in a recent work [6]. In this Letter, we use two simple lattice models, one with implicit and the other with explicit solvent: we show that in the latter the solvent effects cannot be replaced by any effective AA/AA interactions and we propose that this could be the main reason why it is not possible to find global effective interactions when working on real proteins.

The driving force of the folding of globular proteins is the hydrophobic effect [8]. Hydrophobic residues, such as leucine or valine, try to hide their surfaces from the solvent, creating a hydrophobic core surrounded by polar residues. A simple model taking into account this property is the HP model [9–11], where the AA alphabet is reduced to only two species: hydrophobic (H) and polar (P). The hydrophobic effect is captured by an attractive interaction between hydrophobic residues, and the ϵ_{ij} values of the 2×2 interaction matrix are therefore chosen so that $\epsilon_{\mathrm{HH}} < \epsilon_{\mathrm{HP}} (= \epsilon_{\mathrm{PH}}) < \epsilon_{\mathrm{PP}}$. The set of possible conformations is also reduced, describing proteins as self-avoiding walks (SAWs) on a d-dimensional lattice. The energy of a given sequence S mounted on a conformation Γ becomes:

$$H(S, \Gamma) = \sum_{i<j} \epsilon^{nn}_{p_i p_j} \Delta_{nn}(\vec{r}_i, \vec{r}_j), \tag{1}$$

where $\epsilon^{nn}_{p_i p_j}$ are the interaction potentials described above ($p_i = \mathrm{H}, \mathrm{P}$) and $\Delta_{nn}(\vec{r}_i, \vec{r}_j)$ is a contact matrix defined as:

$$\Delta_{nn}(\vec{r}_i, \vec{r}_j) = \begin{cases} 1 & \text{if } i, j \text{ are nearest neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This model has been thoroughly investigated and, despite the strong simplifications, it has provided many results in agreement with the properties of real proteins, such as the presence of a hydrophobic core in the native state of proteins at low temperature and the occurrence of heat denaturation. To check the effects of an explicit solvent we use the HPW (HP + water) model [12]: as in the HP model, proteins are described as SAWs on a lattice and are made of H and P residues. The difference resides in the addition of the solvent: sites not occupied by an AA are occupied by groups of water molecules. The water behavior is described by the Muller-Lee-Graziano (MLG) model [13]. In this water model, as in real water, hydrogen bonds between molecules can be either formed or broken. The typical energies (E_{ij}) and degeneracies (q_{ij}) of these two states depend on whether the hydrogen bond is close to a hydrophobic molecule or in the water bulk. Water molecules close to a hydrophobic AA try to build icelike cages, formed by strong hydrogen bonds, whereas in the bulk, they are able to form hydrogen bond networks. Such a double bimodal description of water is represented pictorially in Fig. 1. Both experiments and simulations suggest that $E_{ds} > E_{db} > E_{ob} > E_{os}$ as from Fig. 1 and that $q_{ds} > q_{db} > q_{ob} > q_{os}$ (subscripts: $d =$ disordered, $o =$ ordered; $b =$ bulk, $s =$ shell; shell sites are those in contact with H AA) [13]. The energy of a protein of L AA, with the sequence $S = p_1, p_2, \ldots, p_L$ ($p_i = \mathrm{H}, \mathrm{P}$) is then:

$$E = \sum_{\langle i,H\rangle}[E_{os}\tilde{\delta}_{i,os} + E_{ds}(1 - \tilde{\delta}_{i,os})] + \sum_{(j,H)}[E_{ob}\tilde{\delta}_{j,ob} + E_{db}(1 - \tilde{\delta}_{j,ob})], \qquad (3)$$

where the first sum is over the water sites that are nearest neighbors (nn) of at least one H AA and the second is over all the bulk sites; $\tilde{\delta}_{i,os} = 1$ if the shell site i is in one of the q_{os} ordered states, 0 otherwise, and analogously $\tilde{\delta}_{i,ob}$ for the bulk sites. The energy leads to the partition function of the system: $Z(S) = \sum_\Gamma Z(S, \Gamma)$, where $Z(S, \Gamma)$ is the partition function associated to a single conformation Γ

$$Z(S, \Gamma) = (q_{ob}e^{-\beta E_{ob}} + q_{db}e^{-\beta E_{db}})^{n_b(\Gamma)}(q_{os}e^{-\beta E_{os}} + q_{ds}e^{-\beta E_{ds}})^{n_s(\Gamma)}, \qquad (4)$$

where $n_s(\Gamma)$ is the number of water sites nn of H residues and $n_b(\Gamma)$ the number of bulk water sites.

To determine the native state of a sequence, we need to introduce a cost function $X(S, \Gamma_i)$ so that Γ_n will be the native conformation of a given sequence S, if

$$X(S, \Gamma_n) < X(S, \Gamma) \quad \forall\Gamma \neq \Gamma_n. \qquad (5)$$

This condition is crucial, but not trivially satisfied, since the conformation with the lowest cost is often degenerate. To avoid this degeneracy we need to create a set $\{S_i\}_L$ of "good" sequences: those with a unique native state. Using the partial free energy

$$X(S, \Gamma_i) = -k_B T \ln[Z(S, \Gamma_i)/Z(S)] \qquad (6)$$

as cost function, we find $\{S_i\}_L$ by exact and exhaustive enumeration: all possible sequences of length L have to be mounted on all conformations and the M_L sequences that satisfy (5) are retained (e.g., $M_{16} = 2431$) [12,14]. The features of these sequences were thoroughly investigated and some typical behaviors of real proteins were reproduced: for instance, the presence of a hydrophobic core in the native state and the occurrence of both cold and warm denaturations. It is worth stressing that cold denaturation is not reproduced by most Hamiltonians used in protein folding, despite the experimental evidence of this phenomenon on real proteins [15].

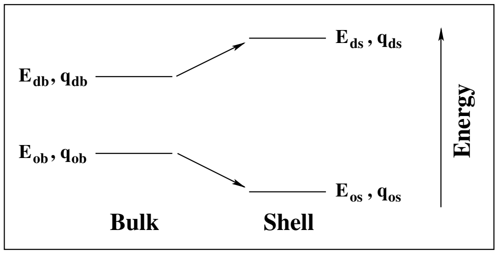
E_{ds}, q_{ds}
E_{db}, q_{db}
E_{ob}, q_{ob}
E_{os}, q_{os}
Energy
Bulk
Shell

FIG. 1. The MLG water model. Bimodal energy distributions for bulk and shell water molecules. The lower levels represent ordered groups of water molecules; the higher levels disordered ones.

The goal is now to find effective AA/AA interactions able to recover the original native states. The general form of a pairwise contact nn Hamiltonian is given by (1). Introducing the notation

$$\epsilon_1 = \epsilon_{\mathrm{HH}}, \qquad \epsilon_2 = \epsilon_{\mathrm{HP}} = \epsilon_{\mathrm{PH}}, \qquad \epsilon_3 = \epsilon_{\mathrm{PP}}, \qquad (7)$$

we can rewrite (1), for a sequence S on a conformation Γ, as:

$$H(S, \Gamma) = \sum_{i=1}^{3} \epsilon_i C_i(S, \Gamma), \qquad (8)$$

where $C_i(S, \Gamma)$ is the number of nn contacts of type i. If the ϵ_i's values are correct, the native conformation Γ_n will have the lowest Hamiltonian value:

$$H(S, \Gamma) - H(S, \Gamma_n) > 0 \quad \forall\Gamma \neq \Gamma_n. \qquad (9)$$

Using (8), inequalities (9) can be rewritten as:

$$\sum_{i=1}^{3} \epsilon_i C_i(S, \Gamma) - \sum_{i=1}^{3} \epsilon_i C_i(S, \Gamma_n) = \vec{\epsilon} \cdot \vec{C}^{\Delta}(S, \Gamma, \Gamma_n) > 0 \quad \forall\Gamma \neq \Gamma_n, \qquad (10)$$

where we introduced the quantity $C_i^{\Delta}(S, \Gamma, \Gamma_n) = C_i(S, \Gamma) - C_i(S, \Gamma_n)$: the difference of the number of contacts of kind i between conformation Γ and Γ_n. Finally, since the interaction potentials should not depend on the chosen sequence, (10) becomes:

$$\vec{\epsilon} \cdot \vec{C}^{\Delta}[S, \Gamma, \Gamma_n(S)] > 0 \quad \forall S = 1, \ldots, M_L \quad \forall\Gamma \neq \Gamma_n(S). \qquad (11)$$

In these inequalities, the vector $\vec{C}^{\Delta}[S, \Gamma, \Gamma_n(S)]$ depends only on the topology (AA positions) of the conformations Γ and Γ_n, whereas the vector $\vec{\epsilon}$ contains the residue interaction values. If a vector $\vec{\epsilon}$ exists satisfying all inequalities (11), it can be found by PERCEPTRON learning. Starting from a trial vector $\vec{\epsilon}_t$ all scalar products from inequalities (11) are computed and the vector $\vec{C}_l^{\Delta}$ with the lowest value of $\vec{\epsilon}_t \cdot \vec{C}^{\Delta}$ is identified. The trial vector is then updated: $\vec{\epsilon}_{t+1} = \vec{\epsilon}_t + \eta\vec{C}_l^{\Delta}$ (usually $\eta \ll 1$), the inequalities are reevaluated with the new vector $\vec{\epsilon}_{t+1}$, and the cycle is repeated [16]. If a solution exists, i.e., a vector $\vec{\epsilon}_c$ exists satisfying all inequalities (11), the vector $\vec{\epsilon}_t$ converges toward $\vec{\epsilon}_c$ and the value $\vec{C}_l^{\Delta} \cdot \vec{\epsilon}_t$ becomes positive in a finite number of steps [5]; otherwise, the algorithm runs forever. The absence of a solution implies that the parametrization of the Hamiltonian is wrong.

First, we apply the method to the standard HP model in 2D, investigating sequences of length $L = 16$. Protein design on this model gives us the database with the native states $\Gamma_n(S_i)$ and the excited conformations $\Gamma(S_i)$ (henceforth called decoys). Since, for a given sequence, the excited states are degenerate, we randomly choose, for each sequence in $\{S_i\}_{16}$, only 15 decoys from the first three states (5 decoys per state). Restricting the choice on the three first states is reasonable since competition takes place mainly between the native state and similar conformations, i.e., those in the first few excited states. The vectors $\vec{C}^{\Delta}$ in (11) are easily computed using the contact matrix (2) and starting from a random vector $\vec{\epsilon}_t$ we iterate the learning process. As expected the perceptron recovers the correct interaction values (i.e., $\epsilon_{ij}^{\mathrm{perc}} \cong \epsilon_{ij}$ with $i, j =$ H, P): if the original interactions that stabilize the native state are only AA/AA interactions, the learning algorithm is able to find them.

To apply the method to the HPW model we use different kinds of Hamiltonians, since we do not know, *a priori,* what kind of contacts would be able to replace the solvent effects. Keeping the same structure of (1), we introduce a new energy function:

$$H(S, \Gamma) = \sum_{i<j} \epsilon_{p_i p_j}^{nn} \Delta_{nn}(\vec{r}_i, \vec{r}_j) + \sum_{i<j} \epsilon_{p_i p_j}^{nnn} \Delta_{nnn}(\vec{r}_i, \vec{r}_j) + \sum_{i<j<k} \epsilon_{p_i p_j p_k}^{3B} \Delta_{3B}(\vec{r}_i, \vec{r}_j, \vec{r}_k), \tag{12}$$

with the contact matrices:

$$\Delta_{nn}(\vec{r}_i, \vec{r}_j) = \begin{cases} 1 & \text{if } i, j \text{ are } nn, \\ 0 & \text{otherwise} \end{cases},$$

$$\Delta_{nnn}(\vec{r}_i, \vec{r}_j) = \begin{cases} 1 & \text{if } i, j \text{ are } nnn, \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

$$\Delta_{3B}(\vec{r}_i, \vec{r}_j, \vec{r}_k) = \begin{cases} 1 & \text{if } i, j, k \text{ are } 2nn \text{ and } 1nnn, \\ 0 & \text{otherwise} \end{cases},$$

where $nnn =$ next nearest neighbors. In this work we use three different Hamiltonians: *nn* contacts (fixing $\epsilon_{ij}^{nnn} = \epsilon_{ijk}^{3B} = 0$), $nn + nnn$ contacts ($\epsilon_{ijk}^{3B} = 0$) and many-body ($nn + nnn + 3B$) contacts. As for the HP model we deal with sequences of length $L = 16$: for each sequence in $\{S_i\}_{16}$, we take 15 decoys, randomly chosen from the first three excited states and we compute the vectors $\vec{C}^{\Delta}$ [17]. Then we investigate the learnability of specific interaction values for small sets of sequences $\{S\}^i \subset \{S\}$. First, we apply the learning algorithm to two randomly chosen sequences. Then, at every new step, if a vector $\vec{\epsilon}_c$ exists satisfying (11), we add a new sequence and we reapply the algorithm. For each run i, if the algorithm does not find a solution in a predefined number of steps N, we retain the size of the group $\{S\}^i$ and the related vector $\vec{\epsilon}_c^{\,i}$. Figure 2 shows the distribution of the sizes of these groups, computed on 1000 runs. For each run i, the size of the group $\{S\}^i$ is represented on the x axis in [%] of the whole set of good sequences ($M_{16} = 2431$) and the distribution of these values is shown on the y axis. As expected, the number of learnable sequences increases if the number of involved interactions increases, too. The number of parameters in the Hamiltonian is too small if we use only *nn* contacts. Indeed, many sequences have at least one excited conformation possessing exactly the same number of contacts as its native state: the latter cannot be recognized as native among the other conformations. This degeneracy is partially avoided by introducing *nnn* contacts, and three-body interactions allow a more accurate tuning between the different potentials: the result becomes more robust and the typical size of a learnable set increases. Still, it is much smaller than M_{16}. The value of the different interactions for the many-body case are shown in Fig. 3. The peaks of the *nn* and the *nnn* values satisfy the inequalities $\epsilon_{\mathrm{HH}} < \epsilon_{\mathrm{HP}} < \epsilon_{\mathrm{PP}}$ as expected from the energy definition of the solvent. Figures 2 and 3 show that the method divides the global set of sequences in small subsets: every subset has its own specific potential values. It is nevertheless important to stress that the algorithm stops because it has not found a solution in N steps, not because the problem has proven to be unlearnable. Therefore, we also apply a modified algorithm, which was first introduced in Ref. [18] and applied to the crambin problem [6]. This algorithm proves unlearnability monitoring a despair parameter d: if this parameter d exceeds a critical value d_c after a finite number of steps, the problem is unlearnable. First, we apply this algorithm using the whole set of sequences for the

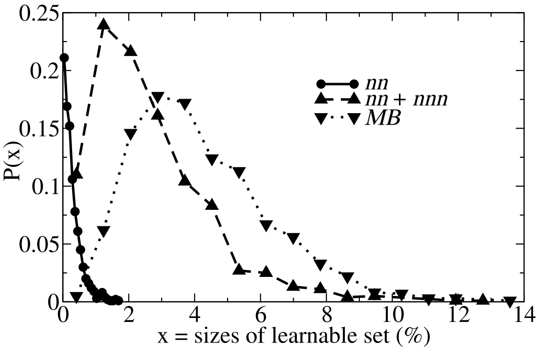

FIG. 2. Probability distribution of the sizes of learnable groups. For each run, the PERCEPTRON finds a set of values satisfying a maximal number of x sequences, represented in [%] of the total on the x axis. On the y axis the probability distribution, computed on 1000 runs, is shown.

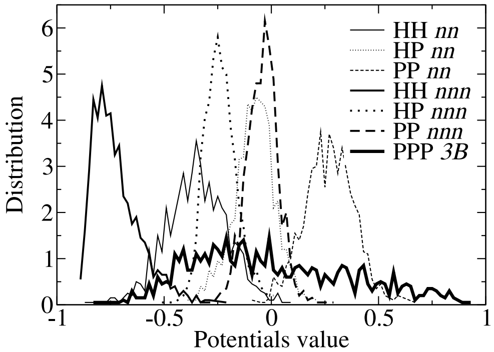

FIG. 3. Probability distribution of the potentials values. (The HHH, HHP, and HPP 3B interactions are peaked at the zero value and for clarity omitted in the picture.)

many-body case. As expected, the result is in agreement with what happens for real proteins: the problem is rigorously unlearnable. Instead, when looking at the small sets $\{S\}^i$, we find that roughly 20% of them are indeed unlearnable, whereas the remaining 80% correspond to the worst case of convergence for the theorem [18], and unlearnability, if any, could be proven only in $N \approx 10^{30}$ steps. Yet, since we know that the whole set is unlearnable, we look at upper bounds for the sizes of the sets $\{S\}^i$: to each of them we add sequences until unlearnability is rigorous. We find that just a few extra sequences are enough and the peak of the histogram in Fig. 2 moves to about 5%.

We further investigate the excited states given by the effective Hamiltonian (12) in the many-body case. Since the common picture of the energy landscape of proteins is that the native state lies at the bottom of a funnel, and folding proceeds along its walls, scrambling the energies of the excited states can affect dramatically the dynamics. Therefore, we recompute, for different sequences S_i, the energies of the conformations in the three first excited states [according to the explicit solvent Hamiltonian (6)], using the effective Hamiltonian (12). For all sequences the initial hierarchy of the excited states imposed by (6) changes if we use (12): for two different conformations Γ_i and Γ_j, the inequality relating the two energies is $E^{\rm HPW}(\Gamma_i) < E^{\rm HPW}(\Gamma_j)$ using (6), but it may be reversed using (12), so that $E^{\rm eff}(\Gamma_i) > E^{\rm eff}(\Gamma_j)$. This energy scrambling may result in a change of the intermediate states, as well as of the folding pathway.

In conclusion, we used a common method to extract interactions potentials on real proteins: the PERCEPTRON method. We applied this algorithm to a simple model where the solvent is explicitly taken into account, showing that the solvent effects cannot be reproduced by effective residue interactions. It turns out that proteins are divided into groups: for each group we find interactions values able to recognize the correct native state of the sequences inside this group. Yet it is not possible to determine global values, meaning that the solvent effects are more complex than simple AA/AA interactions. Moreover, the energy landscapes determined by the new effective Hamiltonians are different from the ones of the solvent dependent Hamiltonian. As a consequence, the dynamical folding process may change, too. These two main results may explain why it is not possible to determine global AA/AA interactions for real proteins and suggest that, even for restricted protein sets, some care should be taken when using effective potentials for dynamical studies.

We thank the Swiss National Science Foundation for support through Grants No. 21-61397.00 and No. 2000-67886.02.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] U. Mayor *et al.*, Nature (London) **421**, 863 (2003).
[3] S. Lifson and C. Sander, Nature (London) **282**, 109 (1979); J. Mol. Biol. **139**, 627 (1980); S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[4] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).
[5] F. Rosenblatt, *Principles of Neurodynamics* (Spartan Books, New York, 1962).
[6] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11 101 (1998).
[7] M. Vendruscolo, R. Najmanovich, and E. Domany, Proteins **38**, 134 (2000).
[8] C. Tanford, Science **200**, 1012 (1978); P. L. Privalov and S. J. Gill, Adv. Protein Chem. **39**, 191 (1988).
[9] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989); H. S. Chan and K. A. Dill, Phys. Today **46**, No. 2, 24 (1993).
[10] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Proteins **32**, 80 (1998).
[11] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).
[12] P. De Los Rios and G. Caldarelli, Phys. Rev. E **62**, 8449 (2000); G. Caldarelli and P. De Los Rios, J. Biol. Phys. **27**, 229 (2001).
[13] B. Lee and G. Graziano, J. Am. Chem. Soc. **118**, 5163 (1996); K. A. T. Silverstein, A. D. J. Haymet, and K. A. Dill, J. Chem. Phys. **111**, 8000 (1999).
[14] G. Salvi, S. Mölbert, and P. De Los Rios, Phys. Rev. E **66**, 061911 (2002).
[15] P. L. Privalov, CRC Crit. Rev. Biochem. Mol. Bio. **25**, 181 (1990).
[16] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).
[17] We checked that we can always learn the parameters for a single sequence even using all non-native conformations as decoys.
[18] D. Nabutovsky and E. Domany, Neural Comput. **3**, 604 (1991).