Design of lattice proteins with explicit solvent

G. Salvi,[1] S. Mölbert,[1] and P. De Los Rios[1,2]
[1]*Institut de Physique Théorique, Université de Lausanne, CH-1015, Lausanne, Switzerland*
[2]*INFM, Sezione di Torino-Politecnico, Corso Duca degli Abruzzi 24, 10100 Torino, Italy*
(Received 22 May 2002; published 30 December 2002)

Protein design is important to develop new drugs. As such, a knowledge of the correct model to use to design novel proteins is of the utmost importance. Here we show that a simple model where the solvent degrees of freedom are (semi)explicitly taken into account performs better than other existing models when compared to real data. Some consequences on the criteria to be used for protein design are discussed.

DOI: 10.1103/PhysRevE.66.061911 PACS number(s): 87.15.By, 87.14.Ee, 36.20.Ey, 87.15.Aa

I. INTRODUCTION

Protein folding stands as one of the major interdisciplinary challenges of the last 10 years, involving biology, chemistry, medicine, and physics. Not less important, especially for the development of new drugs, is protein design [1]. In protein design one chooses a geometrical conformation, and searches some of the amino-acid (a.a.) sequences that have it as a native state. The rationale behind this scheme is that a protein designed to interact with a particular site of a target should have a given shape, to be geometrically compatible, and suitable amino acids on the surface to provide the right chemical properties.

To decide whether a sequence S has its native state on a target structure Γ, one has to be able to give a cost C to all possible structures when mounting S on them. Then, Γ is the native state of S if $C(S,\Gamma)<C(S,\Gamma')$, for any structure $\Gamma' \neq \Gamma$. A suitable cost function, and the most intuitive one, could be one of the Hamiltonians in use for protein folding. Such Hamiltonians are usually a.a./a.a. contact interactions of some kind, encoded in 20×20 matrices (there are 20 naturally occurring amino acids of relevance for protein synthesis, with just a few exceptions) [2]. Although 20×20 interaction matrices are useful for real protein design, they are too complex to address general questions. Still, proteins can be designed with simpler Hamiltonians such as the *HP* Hamiltonian [1,3]. There, only two species of a.a. are defined, either hydrophobic (H) or polar (P), and, in its simpler form, the 2×2 interaction matrix has elements $\epsilon_{HH}=-1$, $\epsilon_{HP}=\epsilon_{PH}=\epsilon_{PP}=0$. When working on a two-dimensional lattice the simplifications introduced in both the a.a. alphabet and in the set of possible configurations allow for *exhaustive* design for proteins of length up to 20 a.a. even on a desktop PC. Although exhaustivity is not the goal of protein design, it can provide insight in the statistical properties of proteins.

Among the relevant statistical features, there are of course thermodynamic quantities. In particular, it is known that the free energy difference between unfolded states and the native state is positive below the warm denaturation temperature T_w, ensuring the stability of the native state; yet, this difference has a maximum around 20 °C for the ''average'' protein, and then decreases for temperatures $T<20$ °C. Either choosing suitable proteins, or, more generally, by supercooling or by applying a pressure, it is even possible to see the *cold* denaturation of proteins in liquid water at a temperature T_c that is, usually, below 0 °C [4]. This phenomenology is not reproduced by most Hamiltonians used in protein folding, since their native state is the $T=0$ ground state of the model. The maximum, at intermediate temperatures, of the free energy and, eventually, cold denaturation, could be recovered by introducing some dependency of the interactions on temperature, but it would be quite an arbitrary one if not derived from some microscopic model.

The paper is organized as follows: in Sec. II we describe a protein model where the solvent is (semi)explicitly taken into account; in Sec. III we present the numerical results of the exact enumeration; Sec. IV is devoted to a discussion of the stability criteria for designable proteins and finally, in Sec. V, we try to draw some conclusions.

II. MODEL DESCRIPTION

To design proteins we use a protein-solvent simplified model that has been recently introduced to describe cold and warm denaturation within the same framework [5]. The main point of the model is that hydrogen bonds in liquid water can be either formed or broken. The typical energies and degeneracies of these two states depend on whether the hydrogen bond is close to a nonpolar (hydrophobic) molecule or in the bulk of water. Such a double bimodal description of water (better represented pictorially, see Fig. 1) has been recently introduced to fit experiments of solvation and rederived by both simple and realistic models of water [6]. 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

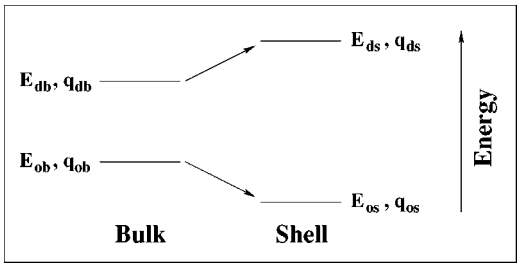

FIG. 1. Bimodal energy distributions for bulk and shell water molecules. The lower levels represent ordered groups of water molecules, the higher levels disordered ones.

those in contact with H a.a.). This explicit (or semiexplicit) description of the solvent allows us to introduce a protein model where there is no need to introduce effective solvent-mediated interactions, nor arbitrary temperature dependencies.

Proteins are described as self-avoiding walks (SAW) on a two-dimensional square lattice. Sites not occupied by amino acids are occupied by groups of water molecules (here we introduce some coarse graining in the model): the water degrees of freedom are represented by Potts-like variables σ that take $q_{os}+q_{ds}$ ($q_{ob}+q_{db}$) values for shell (bulk) sites. Given a protein of L amino acids, with the sequence $S=a_1,a_2,\ldots,a_L$ ($a_i=P$ or H), the energy of the protein is then

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

where the first sum is over the water sites that are nearest neighbors of some H a.a. and the second is over all the bulk sites; $\tilde{\delta}_{i,os}=1$ if $\sigma_i=0,\ldots,q_{os}-1$, 0 otherwise, and analogously for $\tilde{\delta}_{i,ob}$. Due to the similarity with the *HP* model and to the explicit presence of water, we refer to this model as the *HPW* model (*W* stands for water). Starting from Eq. (1) we can write the partition function of the system as $Z(S)=\Sigma_\Gamma Z(S,\Gamma)$, where $Z(S,\Gamma)$ is the partition function (which can be considered as a *generalized* Boltzmann weight) associated to a single conformation Γ,

$$Z(S,\Gamma)=(q_{ob}e^{-\beta E_{ob}}+q_{db}e^{-\beta E_{db}})^{n_b(\Gamma)}\times(q_{os}e^{-\beta E_{os}}+q_{ds}e^{-\beta E_{ds}})^{n_s(\Gamma)}, \tag{2}$$

where $n_s(\Gamma)$ is the number of water sites nearest neighbors of some H a.a. and $n_b(\Gamma)$ is the number of bulk water sites. From the partition function we can calculate all the thermodynamic quantities. In particular, the specific heat and the thermal average $\langle n_s\rangle$ as a function of T point to two different denaturations, since the C_v shape contains two well-defined peaks (see Fig. 2; details are given in the following section). Between the two peaks proteins have few contacts with water (they are in a compact state), and the most probable conformation is the one with the minimum $n_s(\Gamma)$ contacts (H a.a. are hidden in the core of the protein, as for real globular proteins). Above and below the two temperatures proteins swell and the number of water-protein contacts increases. So, the *HPW* model captures within a single framework both warm and cold denaturations.

III. NUMERICAL RESULTS

We tackle the design problem by exact enumeration. The cost function of sequence S mounted on structure Γ is $C(S,\Gamma)=-Tk_B\ln[Z(S,\Gamma)/Z(S)]$, that is the partial free energy of configuration Γ. Then it is easily seen that the smaller $n_s(\Gamma)$, the lower $C(S,\Gamma)$. Given the set $\{S_L\}$ of all the possible sequences of length L, and the set $\{\Gamma_L\}$ of all the possible conformations, we look for those sequences $\{S'_L\}$ that have a unique state with a minimum $n_s(\Gamma)$ among all the possible conformations. We call $\{\Gamma'_L\}$ the corresponding set of native structures. The uniqueness of the native state is a requirement to ensure that the folding to the correct conformation is not hindered by the competition of different, thermodynamically equivalent, states. Increasing the length of the proteins from 18 to 20 forbids us to still use exhaustive enumeration due to the exponential growth of the computation time. Yet, we can use the information about the compactness of the native state to reduce the number of configurations that we should check. Indeed, we find that all the sequences $\{S'_L\}$ ($L\leqslant 18$) have native states $\{\Gamma'_L\}$ with perimeter limited by some $p_{max,L}$; tentatively, we use only conformations with $p\leqslant 21$ for the exact enumeration for $L=20$. We find $N_{S,20}=37\,933$ sequences that are good proteins, finding their native states on $N_{\Gamma,20}=5440$ with perimeter up to $p_{max,20}=20$. To check if we have left apart some good proteins, we tried to extrapolate the number of sequences in $\{S'_{20}\}$ using the data for $L=10,\ldots,18$. Considering that the two related sets of walks on a lattice, i.e., the Hamiltonian walks and the self-avoiding walks, have both an exponential growth, we used an exponential fit for our data. The extrapolated curve overshoots $N_{S,20}$ by 1%, which is compatible with the approximation of the curve for $L\leqslant 18$ (Fig. 3). We can conclude that, even if in principle there could be other sequences in $\{S'_{20}\}$ that we did not find or degenerate competitors that we did not consider, still they should not represent a significative modification of the set. The number of designable conformations $\{\Gamma'_L\}$ as a function of their length grows with a connective constant $\mu_{des}\simeq 1.74$, close to $\mu_{HW}\simeq 1.47$ typical of Hamiltonian walks (SAWs have $\mu_{SAW}\simeq 2.63$), which, together with the perimeter data, confirms that native states are compact.

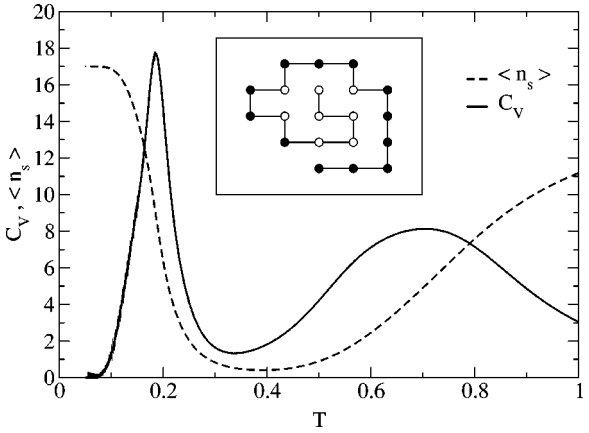

FIG. 2. Specific heat C_V (out of scale) and $\langle n_s\rangle$ for the protein shown in the inset; $q_{ob}=500$, $q_{db}=2000$, $q_{os}=1$, $q_{ds}=10000$, $E_{db}=-E_{ob}=1$, $E_{ds}=-E_{os}=2$.

Thermodynamic quantities for the sequences in $\{S'_L\}$ are easily computed. The parameter values of the model were chosen considering some qualitative criteria. We used large numbers for the degeneracy values q_{ob}, q_{db}, and q_{ds} ($q_{os}=1$ being fixed, since the absolute multiplicity is irrelevant). This choice is reasonable since every site contains some water molecules, so that the total number of states per site will

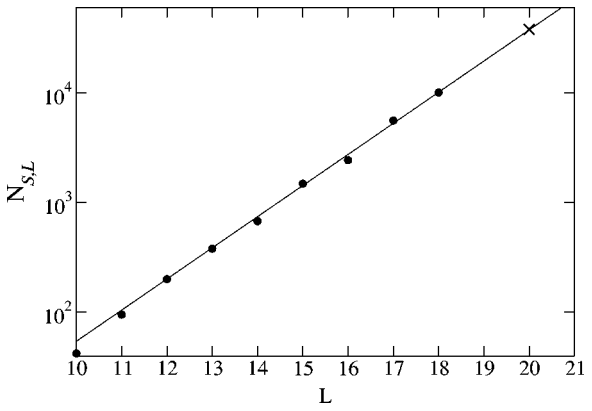

FIG. 3. Number of sequences $N_{S,L}$ (the y axis is logarithmic). The black dots represent the values found by exact enumeration for sequences of length $L=10,\ldots,18$. The line is the extrapolated curve, whereas the cross is the number of sequences in the set $\{S'_{20}\}$, created checking only the configurations with perimeter $p \leqslant 21$.

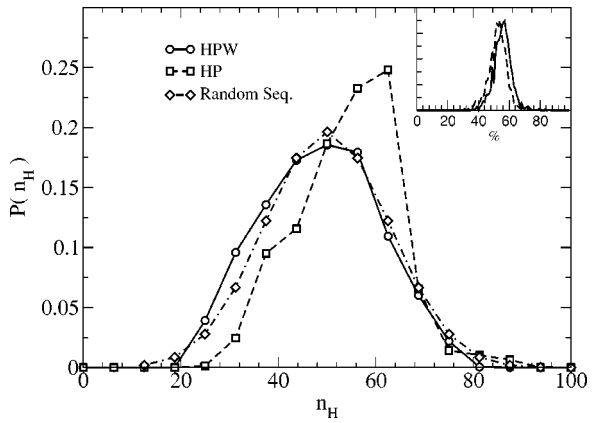

FIG. 4. Percentage n_H of H monomers of the sequences of length $L=16$. The inset shows the result of the FSSP database sequences, using the Eisenberg and Cornette hydrophobic scales.

be the number of states of *one* single hydrogen bond to a power that is the *total* number of hydrogen bonds present in the site. On the other hand, since the energy parameters have no units, we fixed the latters with a symmetry criterion, in order to decrease their degrees of freedom, so that $E_{db}=-E_{ob}=E$ and $E_{ds}=-E_{os}=\eta E$. A better determination of these values could come from molecular dynamics and structural studies, but anyway the results are rather robust to changes of the parameters, as already pointed out in the literature [5]. The results do not change significatively by changing the energy parameters (with and without the symmetry constraint); moreover, they are robust in a range of various order of magnitude of the degeneracy parameters (which is reasonable since a large change in the q's can correspond to even a slight change in the number of states for a single H-bond, due to the power raising).

In Fig. 2, already described above, we have chosen the particular sequence $PPPPPPHPPPHPPHPHHHHH$ and the parameter values $q_{ob}=500$, $q_{db}=2000$, $q_{os}=1$, $q_{ds}=10\,000$, $E_{db}=-E_{ob}=1$, $E_{ds}=-E_{os}=2$. We have set the Boltzmann constant $k_B=1$. We tried various other sequences and lengths, and the results are always qualitatively the same, with slight changes in the peak height and width, and with some small T_w and T_c variations.

As already known from real data, and compatibly with protein design according to other models, we also find that not all the compact structures have the same designability, with some of them more designable than others, in qualitative agreement with the *superfold* (or *fold families*) concept: many proteins with different sequences share the same native fold. Indeed, we find that 62% of all sequences in $\{S'_{20}\}$ have their native state on just the 17% of all designable configurations (those that are native states of 11 sequences or more; the highest designable structure attracts 147 sequences): most proteins find their native fold on a restricted number of structures. Moreover, as the protein length L grows, the number of sequences per native structure increases: the ratio $(N_{S,L}/N_{\Gamma,L})$ is proportional to μ_r^L, with $\mu_r \simeq 1.12$. This happens because, for larger values of L, high designable configurations are present, reducing the ratio of needed structures, and creating larger fold families.

Design according to some model should be tested against as much of the known protein phenomenology as possible, at least qualitatively. *HPW* proteins already recover the correct thermodynamics, compactness, and structure (segregation of a hydrophobic core) of real proteins. Some more information comes from sequence statistics. We look at two basic indicators, namely, the H amino acid concentration and the so-called *run test*. In Fig. 4, in the inset, we have reproduced the hydrophobic content of a sample of proteins chosen from the FSSP database (*fold classification based on structure-structure alignment of proteins* [7]). The two curves correspond to the Eisenberg and Cornette scales [8]; they are well approximated by binomial distributions with mean values around 55–60 %. Since according to these scales the number of hydrophobic a.a. species is 11 and 12, respectively (55% and 60%), binomials peaked around those values are typical of random sequences. For a two-letter alphabet (P and H) random sequences are peaked around 50%. *HP* proteins, instead, are peaked around 60%. On the other hand, the distribution of *HPW* sequences is nicely approximated by a binomial around 50%: the *HPW* model does not introduce any bias toward more hydrophobic sequences then pure chance would do, in apparent similarity with real proteins. The second test we mentioned, the run test, has been introduced in the context of proteins by White and Jacobs [9]. Sequences are reduced to binary strings of Hs and Ps. Then, every series of consecutive Hs (or Ps) is counted as a *run*. As an example, the string *HHHPP* contains two runs (*HHH* and *PP*), and the string *HPHPH* contains five runs (each single letter counting as a run). The run test analyzes the distribution of proteins according to the number of runs they contain. In Ref. [9] it was found that, according to the run test, real proteins are statistically indistinguishable from random sequences. Although we do not want to dwell into the implications of this result, it is important to stress that this is a part of the phenomenology of real proteins and as such it would be auspiciable to recover it through models. In Fig. 5 we show the run distribution for proteins of length $L=16$ de-

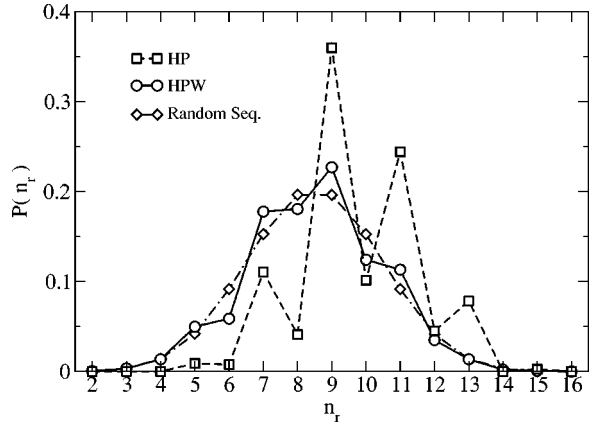

FIG. 5. Number of runs n_r of the native sequences with length $L=16$.

signed using both the *HP* and *HPW* models. Again, *HP* proteins are different from random sequences, with a preference for sequences with many runs. On the other hand, *HPW* proteins again behave similarly to real proteins (and to random sequences). The two sets of data together (*H* percentage and run test) allow us to give an interpretation of the HP behavior: sequences with a high *P* density have a high probability of having long *P* runs. Since in the *HP* model that we considered there are no *HP* and *PP* interactions, these long *P* runs would be free to fluctuate, giving rise to degenerate native states. Therefore, the resulting designable sequences have fewer *P*s and many runs, so to lock the *P*s [10]. The *HPW* model, instead, does not need such a selection since, at an effective level, it introduces also *HP* interactions that forbid large fluctuations of any long *P* run. To check that this is indeed the case, we also designed proteins with a modification of the *HP* model (that we call here $HP2$), with interactions $\epsilon_{HH}=-2.0$, $\epsilon_{HP}=-1.0$, and $\epsilon_{PP}=0$ (we chose $\epsilon_{HH}=2\epsilon_{HP}$ to simulate the degeneracy of the effective interactions coming out of the *HPW* model). Indeed, we find that $HP2$ proteins perform better than *HP* ones on both statistical tests, confirming the above interpretation. Anyway, only about 50% of the $HP2$ proteins correspond to *HPW* proteins and vice versa. This is due to the fact that the *HPW* model introduces a richer effective interaction hierarchy than simple two-body interactions independent of the context (the relevance of many-body interaction terms, and of some context dependence, has been recently pointed out [11–14]).

IV. OVERALL STABILITY CRITERION

Proteins belonging to the $\{S_L'\}$ set have been selected to satisfy the criterion of uniqueness of the native state. Yet, it is not the only request that has to be imposed on proteins. Indeed, the native state has to be stable against other states: at equilibrium it should be the most favorable one, meaning that the probability p_s to occupy it, between the two transition temperatures T_c and T_w, should be at least larger than 1/2 (1/2 being the value at the denaturation transitions). Not all the sequences in $\{S_L'\}$ satisfy this criterion, and a second set $\{S_L''\}$ has to be generated by selecting all those sequences $s\in\{S_L'\}$ such that $p_s>0.5$ for $T\in[T_c,T_w]$.

This means that a new definition of *native* state has to be introduced. The Boltzmann weight of the native state $Z(s,\Gamma_{nat})$, for a given sequence $s\in\{S_L'\}$, has to be larger than the sum over all partition functions of the excited states:

$$\left\{ Z(s,\Gamma_{nat},T) > \sum_{\{\Gamma_{exc}\}} Z(s,\Gamma_{exc},T) \right\}\Bigg|_{T\in[T_c;T_w]}. \qquad (3)$$

This is a completely new criterion in protein design; indeed, it is automatically satisfied by *HP* proteins. Once their native (ground) state is guaranteed to be unique, there will always be a temperature below which $p_s>0.5$. In the *HPW* model, instead, the native state, even having the lowest cost function value, can result unfavorable compared to phase space regions of high cumulated probability, usually related with a high degeneracy of the first excited states.

Proteins in $\{S_L''\}$ have slightly sharper transitions than those belonging to $\{S_L'\}$ but that did not satisfy overall stability; they also pass the H percentage test, but not the run test. Indeed, $\{S_L''\}$ proteins have a run distribution that is quite different from the distribution for random sequences. Here we would like to preserve the statistical features of sequences as an indicator of the model validity. Hence, we try to revise the uniqueness and overall stability criteria to understand if it is possible to relax them, so to recover the good sequence statistics.

Protein folding is a dynamical process taking place at some temperature between T_c and T_w. Uniqueness is usually invoked to ensure that the folding is not misled to a target structure different from the native one. Actually, chances are extremely good for correct folding even in the presence of (almost) degenerate competitors, if the basin of attraction (the *funnel*) of the correct native state is much larger than that of the decoy. We can easily envision the extreme case where the native state has a very large basin of attraction, and there is a competitor that is like a golf hole in the free energy landscape. Clearly, this decoy will almost never be found by the dynamics, and this protein could be retained by natural selection, whereas, the same protein, using the strict criterion given by Eq. (3), would be discarded: although the Boltzmann weight of the competitor plus that of the other non-native states can be larger than that of the native state, the competitor's weight should not be taken into account because, dynamically, it will almost never be found. Therefore, Eq. (3), with the exclusion of the competitor in the rhs(right-hand side), could be satisfied. The method used in the present work gives us no information about the shape of the energy landscape, since only thermodynamic quantities are considered. A detailed study of the structure of the configuration space, and of the *dynamical accessibility* of the native state and of its thermodynamic competitors is needed. This could lead to a further definition of *native* state. Sequences with a unique native state could be retained even if Eq. (3) is not satisfied. This happens in the case where the probability to dynamically fold in a thermodynamic competitor is negligible. Such a new criterion would produce a new set $\{S_L'''\}$ of *good* sequences, which might keep the statistical

properties of the set of sequences $\{S'_L\}$ (the one found using only the uniqueness criterion), the latters being in agreement with the ones calculated on real proteins. Work is in progress in this direction, where a dynamical approach to folding and design is needed, the simplest being Monte Carlo (MC) dynamics (which is not, anyway, the real folding dynamics).

V. CONCLUSIONS

We have approached the protein design problem using a model where the solvent degrees of freedom appear explicitly, and we have elucidated some of the problems of protein design, and folding, by a comparison with part of the known phenomenology of proteins: thermodynamics and sequence analysis. It turns out that such reference to real data is useful to discriminate between models. Also the design criteria themselves come under scrutiny, suggesting that a much relevant role in protein folding should be given back to dynamics, and to a careful study of the structure of the phase space (the possible conformations and the way they are connected to each other by the dynamics) [16]. On the other hand, exact enumeration cannot address dynamical issues, neither can it be used above 2D(two dimensions). Indeed, the sequences selected as *good* proteins show an hydrophobic core in their native configuration and polar amino acids on the surface. To reproduce the same property on a 3D lattice we need a reasonable surface or volume ratio of the structures, which cannot be achieved with sequences of length below $L=20$. Moreover, *HPW* compact native states are not, in general, easily embedded in simple volumes, such as squares (2D) or cubes (3D), so that a full SAW enumeration is still necessary. We are presently working on a MC approach to *HPW* protein folding and design that will allow us to look at 3D longer proteins, and to their related issues. 3D longer proteins should also tell us whether sequence statistics, that can surely be used as a phenomenological indicator of a model's validity, is already meaningful in 2D or, on the contrary, looking at 3D will change the picture (the *HP* model in 3D and for longer proteins could satisfy the tests or could retain the 2D pathologies). Design according to the *HPW* model is moreover necessary to build a database to explore the relevance of three- and many-body effective interactions on the physics of protein when the degrees of freedom of the solvent are traced out [15].

ACKNOWLEDGMENTS

We thank the Swiss National Science Foundation for support through Grant No. 21-61397-00 and F. Becca for enlighting discussions.

[1] C. Micheletti, F. Seno, A. Maritan, and J.R. Banavar, Proteins **32**, 80 (1998).
[2] S. Miyazawa and R.L. Jernigan, Macromolecules **18**, 534 (1985).
[3] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, and M.H. Hecht, Science **262**, 1680 (1993).
[4] P.L. Privalov, CRC Crit. Rev. Biochem. **25**, 181 (1990).
[5] P. De Los Rios and G. Caldarelli, Phys. Rev. E **62**, 8449 (2000).
[6] B. Lee and G. Graziano, Acc. Chem. Res. **22**, 5163 (1996); K.A.T. Silverstein, A.D.J. Haymet, and K.A. Dill, J. Chem. Phys. **111**, 8000 (1999).
[7] http://www2.ebi.ac.uk/dali/fssp/
[8] D. Eisenberg *et al.*, J. Mol. Biol. **179**, 125 (1984); J.L. Cornette *et al.*, *ibid.* **196**, 659 (1987).
[9] S.H. White and R.E. Jacobs, J. Mol. Evol. **36**, 79 (1993).
[10] A somewhat related interpretation was obtained by direct inspection of *HP* proteins by C.T. Shih *et al.*, Phys. Rev. Lett. **84**, 386 (2000).
[11] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998).
[12] P.L. San Biagio, D. Bulone, V. Martorana, D.B. Palma-Vittorelli, and M.U. Palma, Eur. Biophys. J. **27**, 183 (1998).
[13] K. Park, M. Vendruscolo, and E. Domany, Proteins **40**, 237 (2000).
[14] A. Kabakçioğlu, I. Kanter, M. Vendruscolo, and E. Domany, Phys. Rev. E **65**, 041904 (2002).
[15] G. Salvi and P. De Los Rios (unpublished).
[16] A. Torcini, R. Livi, and A. Politi, e-print cond-mat/0103270.