A Self-Consistent Knowledge-Based Approach to Protein Design

Andrea Rossi,* Cristian Micheletti,* Flavio Seno,† and Amos Maritan*
*International School for Advanced Studies and INFM, I-34014 Trieste, Italy and the Abdus Salam International Center;
†INFM-Biophysics, Dipartimento "G. Galilei," 35100 Padova, Italy

ABSTRACT A simple and very efficient protein design strategy is proposed by developing some recently introduced theoretical tools which have been successfully applied to exactly solvable protein models. The design approach is implemented by using three amino acid classes and it is based on the minimization of an appropriate energy function. For a given native state the results of the design procedure are compared, through a statistical analysis, with the properties of an ensemble of sequences folding in the same conformation. If the success rate is computed on those sites designed with high confidence, it can be as high as 80%. The method is also able to identify key sites for the folding process: results for 2ci2 and barnase are in very good agreement with experimental results.

INTRODUCTION

Two of the most investigated problems in molecular biology are protein folding and design. Both problems stem from Anfinsen's discovery (Anfinsen, 1973) that the sequence of amino acids of a naturally occurring protein uniquely specifies its thermodynamically stable native structure. The protein folding challenge consists of predicting the native state of a protein from its sequence of amino acids, while in protein design one is concerned with identifying the amino acid sequences folding into a pre-assigned native conformation. The protein design problem asks which and how many amino acid sequences fold into a given native structure. This last issue, having obvious practical and evolutionary significance, has attracted considerable attention and effort from experimentalists and theorists (Pabo, 1983; Quinn et al., 1994; Shakhnovich, 1994; Seno et al., 1996, 1998b; Deutsch and Kurosky, 1996; Morrisey and Shakhnovich, 1996; Dahiyat and Mayo, 1997; Micheletti et al., 1998a,b, 1999c; Street and Mayo, 1999; West et al., 1999; Zou and Saven, 2000). The difficulty of the protein design problem is enormous because, in principle, a rigorous approach (Seno et al., 1996; Micheletti et al., 1999c) would entail a simultaneous exploration of both the family of viable sequences and the family of physical conformations. By doing so, it would be possible to find the sequences having lower free energy in the target structure than in any other conformation. Stated mathematically, to design a target structure Γ, one needs to identify the sequence of amino acids, s, that maximizes the "occupation probability" according to Boltzmann statistics:

$$P_s(\Gamma) = \frac{\exp(-\beta H_s(\Gamma))}{\sum_{\{\Gamma'\}}\exp(-\beta H_s(\Gamma'))} = \frac{\exp(-\beta H_s(\Gamma))}{Z_s} \quad (1)$$

evaluated at a suitable physiological temperature, $1/\beta = k_B T$. $\{\Gamma'\}$ denotes the family of conformations that can house the sequence s, and $H_s(\Gamma')$ is the energy of the sequence in the conformation Γ'. A first obstacle in using Eq. 1 is the difficulty of determining $H_s(\Gamma)$. However, even assuming the correct knowledge of H, it would be impossible to carry out an exhaustive search of the sequence maximizing $P_s(\Gamma)$, due to the computational difficulty of accurately determining Z_s. Several attempts and approximations have been recently proposed to simplify Eq. 1 (Seno et al., 1996; Shakhnovich, 1994; Deutsch and Kurosky, 1996; Morrisey and Shakhnovich, 1996; Seno et al., 1998a; Micheletti et al., 1998a,b; Micheletti et al., 1996; Rossi et al., 2000; Zou and Saven, 2000) and make it tractable at least within a numerical scheme. These attempts range from neglecting (Shakhnovich, 1994) the s-dependence of Z_s to assuming it depends only on the concentration of amino acids (Micheletti et al., 1998a,b; Zou and Saven, 2000) or to using a cumulant (high-temperature) expansion (Deutsch and Kurosky, 1996). A simple and convenient way to test the efficiency of these approximations consists of using models (Micheletti et al., 1999c; Dill et al., 1995) that are amenable to complete enumeration and hence to a rigorous and unbiased check of the design procedure. Several promising results have been obtained in such frameworks showing how the developed theoretical tools have reached a very high degree of reliability (Lau and Dill, 1989; Chan and Dill, 1993; Dill et al., 1995; Micheletti et al., 1999b; Shakhnovich, 1994). However, despite several efforts (Shakhnovich and Gutin, 1993; Sun et al., 1995; Micheletti et al., 1998a), the extension of this machinery to the design of natural proteins has not yet reached maturity. The reasons are mainly two: 1) the difficulty in giving a reasonable functional form of $H_s(\Gamma)$ (Vendruscolo and Domany, 1999); and 2) the impossibility of verifying whether the predicted sequence really folds in the desired conformation, without performing an expensive real experiment.

These two obstacles are absent in simplified lattice models where $H_s(\Gamma)$ is assigned a priori and the exact solution can be rigorously found. In this paper we investigate the degree of accuracy one can reach when designing natural structures (taken from the Protein Data Bank (PDB)) by

Received for publication 31 May 2000 and in final form 18 October 2000.

Address reprint requests to Dr. Andrea Rossi, SISSA/ISAS, via Beirut 4, I-34014 Trieste, Italy. Tel.: 39-040-2240-460; Fax: 39-040-37-87-528; E-mail: rossi@sissa.it.

0006-3495/01/01/480/11 $2.00

using a simple functional form of $H_s(\Gamma)$ and a limited number of classes of amino acids. The unknown parameters defining $H_s(\Gamma)$ are determined with a strategy (Crippen, 1991; Seno et al., 1998b) based on the observation that physical forms of the energy ought to guarantee that any amino acid sequence should recognize its native state as the conformation with minimum energy score and maximum thermodynamic stability. We use such optimized energy functions to design PDB protein conformations by applying some of the above-mentioned theoretical techniques. Finally, we check the quality of our predicted sequences not only through a mere comparison with the naturally folding amino acid sequences (retrieved from the PDB), but performing a statistical analysis of our results with respect to the full set of homologous sequences (e.g., sequences folding to the selected protein or in homologous conformations) (Fersht, 1999). In this way we try to establish which amino acids are important to stabilizing the sequence in the target structure, and we compare these sites with sites important for the folding process, i.e., sites belonging to the folding nucleus (Shakhnovich et al., 1996). Furthermore, we show how it is possible to give a degree of reliability to any design attempt.

The paper is organized as follows: in the next section the schematic representation of protein structures is illustrated, together with the energy functions and the classification of amino acids that have been used. In subsequent sections the new strategy to estimate interaction potentials is derived, the design procedure is explained, and results are discussed and summarized. Technical details are given in the Appendices.

PROTEIN MODELING

Two- and three-body energy functions

As is customary in many numerical approaches to folding and design strategies, we shall also adopt a simplified protein backbone representation that neglects amino acid rotameric degrees of freedom. In fact, we shall use the common coarse-grained model of PDB proteins in which each amino acid unit is represented by a centroid placed on the β-carbon (for glycine the coordinates of the centroid can be estimated by the local geometry of the backbone (Park and Levitt, 1996)). According to this procedure any protein conformation, Γ, obtained by a sequence of N amino acids is specified through the $3N$ Cartesian coordinates:

$$\Gamma \equiv (\vec{r}_1^{\,C_\beta}, \vec{r}_2^{\,C_\beta}, \; \ldots, {}^{C_\beta}_{N}) \, . \tag{2}$$

This simplification is mainly dictated by the necessity to deal only with the main protein degrees of freedom but, as we shall mention, it is also particularly appropriate in design contexts. Furthermore, we shall also partition the 20 types of amino acids into a restricted number of classes. This simplification is not dictated by the numerical convenience of dealing with a restricted sequence space (in fact, the design strategy outlined below can be straightforwardly applied to 20 amino acid classes). Rather, the choice follows from the need to have a sound statistical basis for estimating the free energy contribution of interacting amino acid classes and also from the observation that most amino acids in natural proteins can be substituted without disrupting native folds (Kamtekar et al., 1993). Hence, within the present design scheme we aim at predicting the classes of amino acids designing a given structure. As in Street and Mayo (1999), the putative solution could, in principle, be fine-grained into a 20-amino acid alphabet by using steric packing and solvation constraints.

Finally, the last ingredient of our strategy is the introduction of a suitable (free) energy scoring function. The most popular choice adopted in simplified models is the pairwise interaction form

$$H_s^{(2)}(\Gamma) = \sum_{i<j} \Delta_{ij}^{(2)}(\Gamma) B_2(s_i, s_j) \, , \tag{3}$$

where i, j are the positions along the sequence of the amino acids and the sum is taken over all possible pairs. $B_2(s_i, s_j)$ represents the interaction strength of the amino acid pair s_i and s_j. However, only amino acids that are close enough will interact in a non-negligible way. This is enforced with a suitable weight function, or contact map, $\Delta_{ij}^{(2)}(\Gamma) \equiv f(x = |\vec{r}_i - \vec{r}_j|)$, where:

$$f(x) = \frac{1}{2}\tanh(a_0 - x) + \frac{1}{2} \tag{4}$$

and a_0 is a cutoff value that we choose equal to 8 Å.

In addition to this scoring function in Eq. 3, and to assess possible design improvements, we shall adopt also one including three-body interactions:

$$H_s^{(3)}(\Gamma) = H_s^{(2)}(\Gamma) + \sum_{i<j<k} \Delta_{ijk}^{(3)}(\Gamma) B_3(s_i, s_j, s_k) \, , \tag{5}$$

where $\Delta_{ijk}^{(3)}(\Gamma) \equiv \Delta_{ij}^{(2)}(\Gamma)\Delta_{jk}^{(2)}(\Gamma)\Delta_{ki}^{(2)}(\Gamma)$. The matrix B_3 represents the effective three-body interactions among the different classes of amino acids. Indeed, it has been recently suggested that pairwise energies (Vendruscolo and Domany, 1999) may be unsuitable to describe effective amino acid interactions in proteins. Hence, the introduction of three-body terms might be regarded as the first correction term to Eq. 3 in an expansion scheme where all many-body interactions are included.

Partitioning the 20 amino acids into classes

To estimate the interaction-potential matrices B_2 or B_3 appearing in Eqs. 3 and 5, we introduce a suitable classification of the 20 types of amino acids. In an attempt to go beyond previous studies (Sun et al., 1995; Micheletti et al., 1998a) where the two-letter code was used, we decided to subdivide amino acids into three classes (Table 1).

Although many other subdivisions could be possible, adopting the one followed here has the advantage that, besides clustering amino acids according to their chemical similarities, it creates classes which are almost equally populated. Because the B matrices are symmetric, the number of entries to be determined is 6 and 10 for B_2 and B_3, respectively.

LEARNING THE INTERACTION POTENTIALS

A new theoretical approach

An efficient way to estimate the effective potentials B_2 and B_3 was pioneered by Crippen (Maiorov and Crippen, 1992)

TABLE 1 Three-class partition of amino acids

Hydrophobic	Neutral	Charged
Alanine	Asparagine	Arginine
Isoleucine	Cysteine	Histidine
Leucine	Glutamine	Lysine
Methionine	Glycine	Aspartic acid
Phenylalanine	Serine	Glutamic acid
Proline	Threonine	—
Tryptophan	Tyrosine	—
Valine	—	—

and recently optimized and used (van Mourik et al., 1999; Dima et al., 2000). This scheme aims at finding a set of potentials so that, given a protein sequence s, its native state Γ is recognized as having energy substantially below that of any other equally long conformations Γ' (assumed to be outside the native basin of Γ (Huang et al., 1998)). For a generic energy function $H_s(\Gamma)$ this requires:

$$H_s(\Gamma) < H_s(\Gamma') \tag{6}$$

A key difficulty in turning this idea into a powerful automated scheme is the choice/generation of physically viable decoy structures, Γ'. In many instances the decoys are generated by taking compact "chunks" of suitable length from a bank of proteins (gapless threading). Such decoys may not be physical for certain sequences (for example, due to steric clashes) so that the inequalities (Eq. 6) may enforce rather loose or unrealistic constraints on the extracted potentials.

The first goal in this paper is to propose a strategy to overcome this difficulty. Our idea is based on the fact that the thermodynamic stability requirement, Eq. 6, should be simultaneously satisfied as much as possible for a whole set of conformation Γ_c, which compete significantly with the native state.

This thermodynamic requirement can be accomplished by imposing that

$$H_s(\Gamma) \ll \langle H_s \rangle , \tag{7}$$

where the average $\langle \ldots \rangle$ is carried out over all the set Γ_c. In a more mathematical spirit, Eq. 7 can be derived as follows: Eq. 1 gives the statistical probability that a given sequence s is in a specific conformation Γ at temperature T. If Γ is the native state of s, below the folding temperature only the conformations present in Γ_c give a nonvanishing contribution to Z_s. By writing $Z_s = \exp(\log Z_s)$ and taking the first-order term in its cumulant (high-temperature) expansion, the condition of maximizing $P_s(\Gamma)$ yields Eq. 7.

Due to the linear dependence of the energies H_2 and H_3 on the contact maps (the only factors that contain geometric information about structures), the r.h.s. of Eq. 7 can be re-cast into the following forms:

$$\langle H_s^{(2)} \rangle = \sum_{i<j} \langle \Delta_{ij}^{(2)} \rangle B(s_i, s_j) , \tag{8}$$

and

$$\langle H_s^{(3)} \rangle = \sum_{i<j} \langle \Delta_{ij}^{(2)} \rangle B_2(s_i, s_j) + \sum_{i<j<k} \langle \Delta_{ijk}^{(3)} \rangle B_3(s_i, s_j, s_k) . \tag{9}$$

Notice that both $\langle H_s^{(2)} \rangle$ and $\langle H_s^{(3)} \rangle$ depend on the sequence s and no more on the structure Γ. A detailed technical description of how the averages in Eqs. 8 and 9 are obtained is presented in Appendix 1. To summarize, the functional dependence of $\langle \Delta^{(2)}(i, j) \rangle$ was determined by inspecting its behavior as a function of i, j. The main difficulty was to find a form suitable to represent the behavior of $\langle \Delta^{(2)} \rangle$ for a variety of protein lengths and families. A very satisfactory "collapse" of data from many structures could be obtained by assuming that $\Delta^{(2)}(i, j)$ merely depends on i and j, irrespective of the chain lengths, for $|i - j| < 16$, as shown in Fig. 1.

This is reasonable because the frequency of "local" contacts is not expected to be influenced by the overall protein shape or length. Contacts between residues with sequence separation larger than 16 are rather rare, hence were modeled by assuming a constant frequency of occurrence, $\Delta_2^{(0)}$. The value of $\Delta_2^{(0)}$ is regarded as a free parameter that is to be tuned separately for each protein length so that the average number of overall contacts, $\Sigma_{i,j}\Delta_{i<j}^{(2)}$, matches the number observed in nature. An analogous procedure was followed for the three-body weight function, whose functional form is shown in Fig. 2. For determining the potentials we consider a set of 31 nonredundant proteins listed in Table 2.

Hence, through Eq. 7 and Eqs. 3, 8 (or 5, 9) we obtained one inequality for each protein in the set (that we shall term *training set*). The determination of the potentials, B, was done by using an efficient algorithm, called perceptron, that is guaranteed to provide the best solution for a whole set of inequalities. The method is outlined in Appendix 2. In our case, we have one inequality for each of the training proteins. Clearly, by suitably choosing the B values, it is possible to make each individual inequality arbitrarily large. The perceptron procedure allows finding the best B values that make all inequalities as large as possible simultaneously. There is no guarantee, however, that the inequalities can all be satisfied. Indeed, as a rule of thumb, when the number of inequalities greatly exceeds the number of parameters, no solution can be found if the functional form of it and/or the approximations involved are not satisfactory.

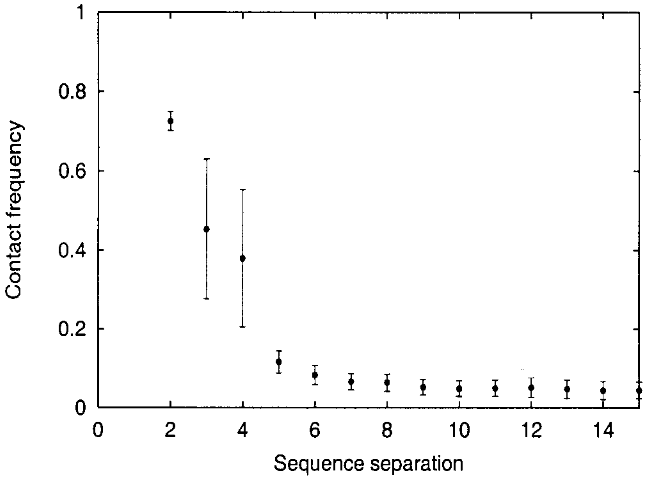

FIGURE 1 $\langle \Delta_{ij}^{(2)} \rangle$ for small values of $k = |i - j|$. For $k = 3$, four long error bars are due to the presence of α and non-α proteins in our protein set.

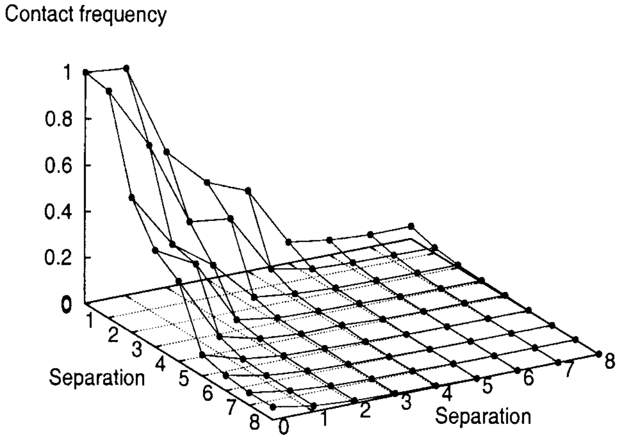

FIGURE 2 $\langle \Delta^{(3)}_{ijk} \rangle$ for small values of $k1 = |i - j|$ and $k2 = |j - k|$. Fluctuations are of the same order as Fig. 1.

In our case we dealt with 5 (or 15) parameters and succeeded in finding physical solutions to the problem. This suggests that the adopted form of the energies was reasonable, otherwise the problem would have been unlearnable. A further proof of this is that, by using a different set of training proteins, nearly the same optimal parameters were obtained, a fact that corroborates the robustness of the potential extraction procedure.

DESIGNING PDB STRUCTURES

The design strategy

Once the potentials are determined, the energy scoring function of any desired conformation can be computed within the energies defined in Eq. 3 or 5. To tackle our ultimate goal, the design of protein conformations, it is necessary to define the design procedure. It has been discussed in the Introduction that a rigorous, but unpractical, way of pursuing this objective consists of finding, for a given conformation Γ^*, the sequence (or sequences) s^* maximizing the occupation probability $P_{s^*}(\Gamma^*)$ defined in Eq. 1. In the previous section we have, however, shown that for the correct energy parameters, the desired sequence should satisfy the inequality:

$$W(s, \Gamma^*) = H_s(\Gamma^*) - \langle H_s \rangle \ll 0\,, \qquad (10)$$

Therefore, since we have obtained a reliable estimate of $\langle H_s \rangle$, we can use Eq. 10 to perform protein design. In practice, given the target conformation, we search for the sequence that minimizes the function $W(s, \Gamma^*)$ where all the quantities are calculated with the above-determined potentials. The optimal solution is identified by a stochastic procedure (simulated annealing) in sequence space, the elementary move being the random mutation of a fraction of residues from one class to another. Generally, the most stringent way to test the reliability/validity of the extracted parameters would be to apply them to design proteins unrelated to the training set. However, as shown in Fig. 12, the extracted potentials varied very little when the training sets 1 or 2 of Table 2 were used (a result that reflects the benefit of the coarse-graining into three amino acid classes). For this reason, to improve statistics on the potentials instead of learning them on set 1 and testing them on set 2, we learned them on the joint set, where the test was carried out.

TABLE 2 List of protein structures

PDB	Length	PDB	Length	PDB	Length
		Set 1			
1acp	77	1beo	98	1cei	94
1coo	81	1cty	107	1erv	105
1fd2	106	1fkb	107	1fna	91
1fow	76	1kum	108	1mit	69
1opd	85	1pdr	99	1rro	108
		Set 2			
1shg	57	1tul	108	1who	96
1yat	113	1yeb	108	2c2c	112
2fxb	81	2imm	114	2mcm	112
2mhr	118	2rhe	114	351c	82
3b5c	93	3ssi	113	3wrp	108
9rnt	104	—		—	

As in Micheletti et al. (1998a) and Sun et al. (1995), the success rate of the design procedure is defined as the fraction of correctly predicted amino acid classes with respect to those of naturally occurring sequences (as found in the PDB) for the chosen configuration. The success rate for a randomly designed sequence where each residue is assigned randomly to one of the three classes would be 33%. For all the considered conformations (see Fig. 3) we obtained a success rate between 40% and 55%.

This success rate can be compared with optimized success rates for two amino acid classes (Micheletti et al., 1998a) which is, on average, ~75%. Clearly, increasing the number of classes makes the problem more difficult, hence a reduced success rate. It is interesting, however, to note that the success rate of the optimal design strategy remains above the random-guessing threshold by ~20%, as for the two-letter case. It is also interesting to notice that this rate does not improve (see Fig. 3) by working with the concentration of amino acid biased toward the composition of the wild-type sequence or even by using the three-body energy. This possibly suggests that important features of real proteins have been equally neglected by all these kinds of energy function.

However, the one-to-one comparison between the designed sequence (defined as the one that minimizes $W(s)$) and naturally occurring ones could not be the best check to do. The reasons are twofold:

- Homologous sequences, e.g., sequences that roughly fold in the same native state, can differ by up to 70% (similarity) of their amino acidic composition. A one-to-one

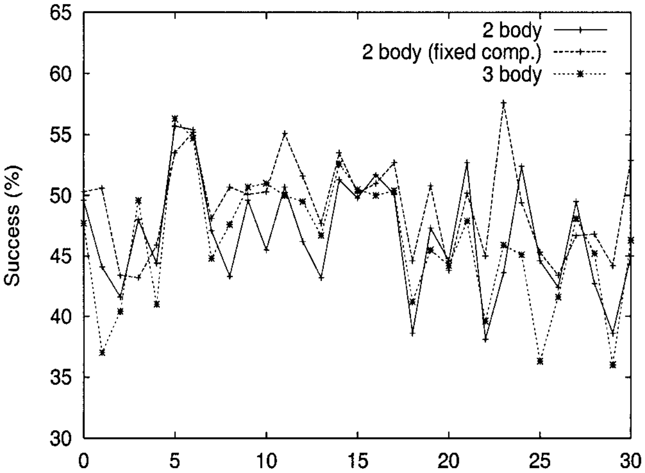

FIGURE 3 The success is here defined as the similarity between the designed sequence and the wild-type sequence as retrieved from the PDB file. The designed sequence has been obtained by a minimization of *W* (simulated annealing) and the success has been obtained as an average over 10 independent minimizations. The three curves refer to the design using Eqs. 3, 7, and 8 with arbitrary or fixed composition, i.e., exploring only sequences with composition not too different with respect to the composition of the wild-type sequence, and using Eqs. 5, 7, and 9.

comparison (although averaged over many sequences) could not be sufficient to verify whether our wrong predictions are involving the most important amino acids or only the marginal ones;

- Naturally occurring proteins may not have necessarily evolved to maximize the occupation probability but also to ensure a fast folding process (Shakhnovich et al., 1996; Maritan et al., 2000a; Hoang and Cieplak, 2000) or maximize uniform compactness (Maritan et al., 2000b). Therefore, to select only the sequences that minimize *W*(*s*) could be a too drastic selection criterion, especially considering that we are working with unperfectly parametrized energy-scoring functions.

To estimate the importance and the effects of these two arguments we performed the analysis discussed below.

Homologous sequences and comparison of similarities

It has been shown by Chothia and Lesk (1986) that naturally occurring sequences with a very low degree of similarity, ~30% (but this rate is very dependent on the length of the alignment (Sander and Schneider, 1991)) can be homologous; that is, they adopt almost the same three-dimensional structure (Fersht, 1999). The original study of Chothia and Lesk was performed using the full repertoire of 20 types of amino acids. In the context of the present study, it is important to estimate how the homology threshold mentioned above changes when the three-letter classification is used. Hence, we re-analyzed the set of protein sequences in the HSSP database (Sander and Schneider, 1991), performing the coarse-graining into H, N, and C classes. The degree of similarity is measured as the percentage of matches between aligned classes rather than individual amino acid types. By definition, the coarse-grained alignment cannot be smaller than the 20-letter one.

The results for a specific protein, 1acp, are given in Fig. 4. It turns out that, on average, the homology threshold of 30% for the full amino acid alphabet corresponds to 55% when the three-letter code is used. This value is remarkably close to the best design scores achieved with our procedure. This does not automatically imply that our solutions are viable. Site-directed mutagenesis experiments have shown that a small fraction of protein sites do not tolerate any substitutive mutation at all (otherwise, the native state would be destabilized). It should then be checked whether such key residues, which are conserved in homologous proteins, are also conserved by our design strategy. In one of following subsections we shall examine this issue in connection with heavily investigated proteins, such as barnase and ci2, and we will show that, as a by-product of the design procedure, the location of such sites can be easily predicted with high reliability. This is not a proof that our design solutions, although different from the native one, are correct, too, but it sheds new light on their validity.

Are extremized sequences the best?

The design analysis we have described so far was based on the selection of sequences that minimize $W(s)$, i.e., on the maximization of the gap between the energy of the sequence in the target conformation and the average energy $\langle H_s \rangle$.

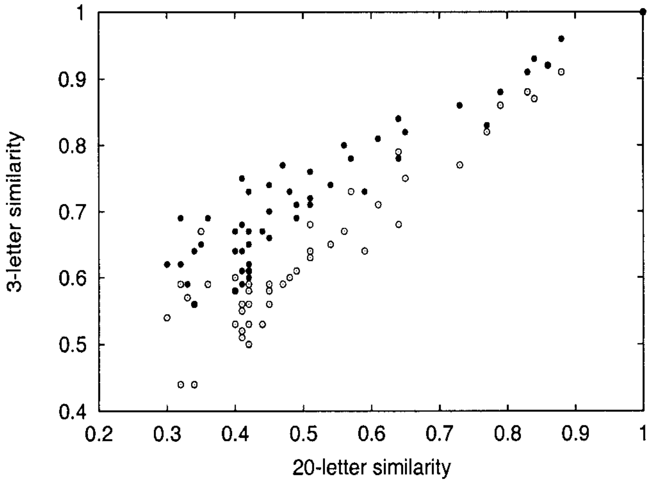

FIGURE 4 Three-letter similarity evaluated for two different classifications versus 20-letter similarity. Filled circles correspond to the classification of amino acids adopted in our design procedure (see Table 1), while open circles correspond to a random repartition of amino acids in classes. The figure refers to a comparison between protein sequence 1acp with 51 sequences of homologous protein.

However, it is presumable that the evolutionary pressure toward rapid and reliable folding (Micheletti et al., 1999a) has not taken the maximization of inequality 10 to the extreme, but to a lower threshold sufficient for biological purposes. For this reason we chose to test the success rate not only for the minimum value of $W(s)$, but also for other sequences. In particular, it is interesting to compare all the sequences s with $W(s) < W(s^*)$, where s^* is the wild-type sequence. For each annealing temperature we extract 100 decorrelated sequences and make a statistical analysis on this sequence set. We evaluate the average of $W(s)$ for this set and a "super-sequence" by applying a pointwise majority rule to this set. In other words, for each site we assign the most frequent amino acid class observed in this sequence set at the given location. Fig. 5 shows the data pertaining to such design attempts on five different proteins. It appears that, indeed, the highest matching with the native sequence is not obtained for the lowest value of W, but for higher ones.

This fact suggests a powerful way to improve the reliability of the design strategy: we can select as putative solutions a wider range of protein sequences and then process the statistical information contained in them to yield a single "super-sequence." Furthermore, one can decide to make a prediction only for those sites where a class has an occurrence frequency larger than some suitable threshold f_0. The number of sites N_s for which we make such a prediction is a decreasing function of f_0, and for a given f_0 depends on the fictitious temperature (at low temperature all the sites are locked). Fig. 6 shows success rates over the N_s betted sites for different values f_0 (data pertain to protein 1erv, other proteins produce analogous plots).

It is evident that when N_s is small, the design procedure is very reliable: retaining the first 40 sites gives the impressive success rate of 80%. It is tempting to conjecture that the residues that are assigned with very little uncertainty by our design procedure (conserved design residues) could also correspond to conserved residues in nature. In the next section we shall examine in detail this possibility, and conclude that there is a significant correlation between the two sets of residues.

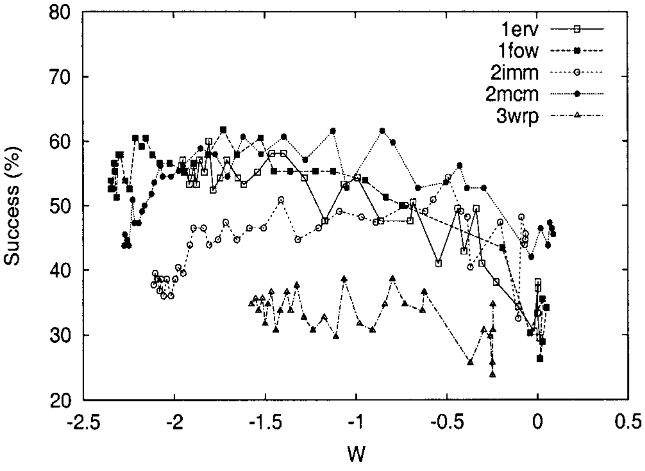

FIGURE 5 The success as a function of the cost function $W(s, \Gamma_t) = H_s(\Gamma_t) - \langle H_s \rangle$ per site. Success is defined here by majority rule on a sampling of 100 (decorrelated) sequences. The value of the cost function for the respective wild-type sequences is between −0.48 and −0.78.

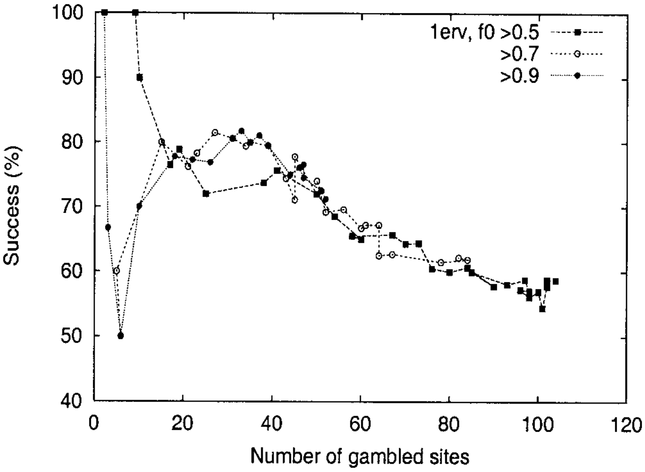

FIGURE 6 Success as a function of the number of betted sites for the protein 1erv. Betting the 40 most locked sites, it is possible to obtain an almost 80% success rate. Note that success is almost independent on the frequency threshold f_0.

Homologous sequences and conserved sites

It is well known (Sander and Schneider, 1991) that homologous sequences present conserved sites, e.g., sites where the type of the amino acid remains unaltered throughout the full set of sequences. In Fig. 7 this fact is graphically elucidated (and even enforced) by analyzing the homologous sequences of protein 1erv with our tripartite classification of amino acids. To each site we assign a color reflecting the conservation of the most frequent class observed in that position. A full conservation of H, N, and C types is denoted with a saturated green, red, and blue color, respectively; the lowest possible conservation of the most frequent class, 1/3, is associated with the white color. According to this scheme, sites with high variability will correspond to lighter nuances.

A visual inspection of the colors assigned to protein 1erv (*top panel* of Fig. 7) reveals that ~30% of the sites are highly conserved. We want to elucidate whether there exists a connection between such conservation of amino acids found in nature and the one emerging in the putative solution obtained from our design procedure.

To do this we performed a simple analysis of the design solutions at different values of the conservation threshold, W. In each batch of 100 design runs, the target value of W was fixed (in a stochastic way) by varying a suitable control

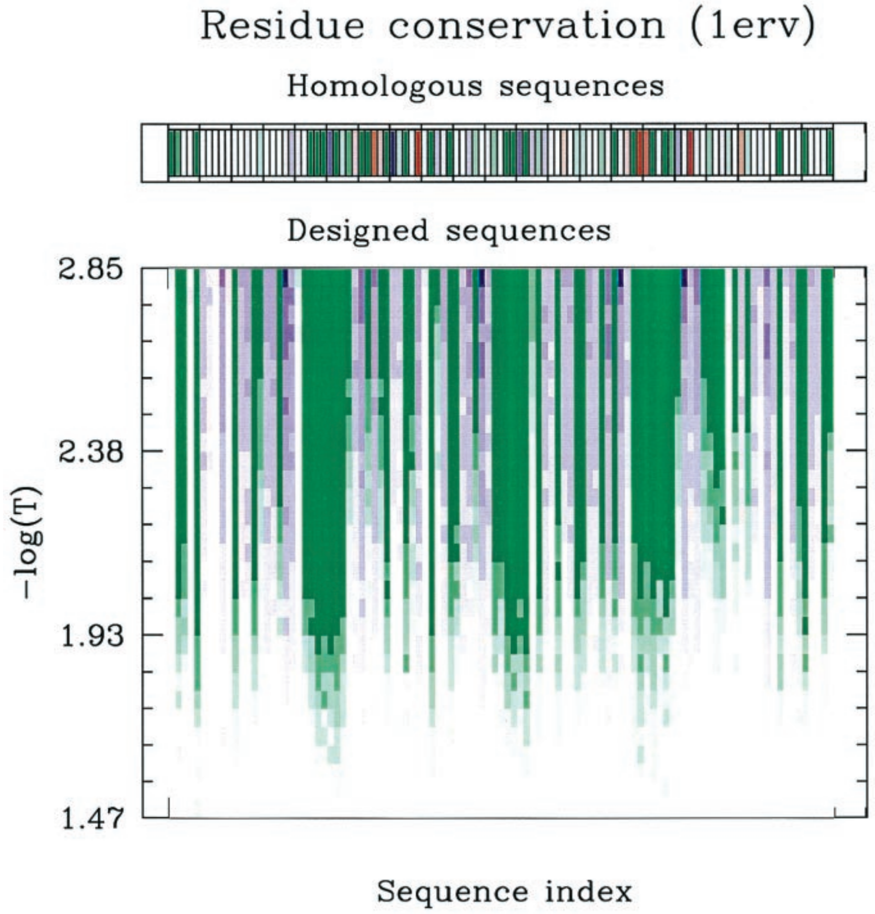

FIGURE 7 Color-coded conservation of residues in protein 1erv (thioredoxin) in natural context (*top*) and in putative solutions obtained with our design procedure. The color code, described in the text, assigns lighter colors to highly variable sites. The conservation in the natural context was obtained from the analysis of the HSSP database (Sander and Schneider, 1991).

parameter, T (by analogy, if we identify W as an energy cost function, T plays the role of the temperature). Finally, for each value of T we analyze the conservation of residues in the designed sequences and color them with the same scheme described above. The results are shown in the large box of Fig. 7.

For high values of T (high W) all the color intensities are very low, indicating a uniform (random) distribution of the classes, but upon decreasing the temperature some of them start to be selected with higher and higher frequency. At very low temperature all the sites are locked in a particular class. This trivial situation is not shown in Fig. 7 which, instead, concentrates on the more relevant range of intermediate temperatures.

The comparison of the native colored panel and the designed one strongly confirms the hypothesis that sites locking early (at high values of T) are related to the naturally conserved ones. This connection is examined in a more circumstantial context in the next section, where we consider two specific protein instances: barnase and chymotrypsin inhibitor.

It is interesting to note that locking occurs first for hydrophobic residues and later for charged ones, a fact reflecting the strength of interactions. Neutral residues, however, appear to have interactions that are relatively small in modulus, and hence contribute much less to the minimization of expression 10. In fact, the locking of neutral residues is observed for temperatures much lower than the ones shown in the plots.

An even more quantitative analysis of the correlation between designed and homologous sequences can be obtained by a simple geometrical construction. For each amino acid located at site i in a given protein, a three-dimensional vector is constructed whose components are the frequencies with which the three classes appear: in the design sequences (we term the vectors $\vec{f}_i^D$) or in the homologous sequences ($\vec{f}_i^N$). To make the comparison meaningful, the design procedure was carried out at a value of T chosen so that the fraction of conserved residues was similar to the one observed in nature. The vector of a site conserved in a specific class of amino acids is aligned with the associated axis, whereas the vector of a nonconserved site has at least two nonvanishing components.

The angle θ_i formed by the two vectors $\vec{f}_i^D$ and $\vec{f}_i^N$ provides a quantitative measure of the correlation between residue conservation in the natural and design contexts. This angle is zero if the agreement is perfect, while it attains the maximum value of $\pi/2 \approx 1.5$ if a residue is maximally conserved in nature and minimally conserved in design (or the other way around).

In Fig. 8 we plot (for four different proteins) the histogram of these correlation angles (*light gray*). Remarkably, for all the proteins the highest entries correspond to small angles, and they represent a considerable fraction (1erv = 24, 2imm = 18, 2ci2 = 12, and 1a2p = 20) of all sites, thus highlighting a highly significant agreement. To validate the design scheme it is then crucial to verify whether the highest agreement (small angles) is observed in correspondence of sites highly conserved in nature. This is indeed the case: in the same figure we plot, for each angle bin, the number of sites that are naturally highly conserved (*dark gray*), i.e., that have a conservation entropy, evaluated as in the HSSP data bank (Sander and Schneider, 1991) lower than ln(1.5) (ln(1) and ln(3) correspond respectively to the minimum and the maximum values for the entropy when only one class is assigned or all three classes are assigned with equal probability). Almost all the sites with a vanishing correlation angle satisfy this property!

We can then conclude that amino acids which, in our design scheme, are designed with a higher confidence, strongly correlate with those that are conserved in natural sequences.

Data for barnase and chymotrypsin inhibitor

In this last section we shall apply the design strategy to two proteins whose folding process has been heavily investigated experimentally. With a series of key measurements (Fersht, 1995; Itzhaki et al., 1995), Fersht and co-workers have identified a restricted set of residues, the folding nucleus, which play a key role in the folding process in proteins such as barnase (1a2p) and chymotrypsin inhibitor (2ci2). Although, generally speaking, naturally occurring proteins can tolerate a fair degree of amino acid substitutions without disrupting the native state, random mutations of sites in the folding nucleus will impair the folding process dramatically. Indeed, recent theoretical studies (Micheletti et al., 1999a) have shown that key sites in the folding process nucleus are part of a bottleneck in the folding kinetic, which is mainly dictated by the native state topology. Overcoming such a bottleneck can occur only through a careful selection of the type of involved amino acids (Cecconi et al., 2000). This novel argument confirms and explains the observation already present in the literature (Shakhnovich et al., 1996) that sites involved in folding nuclei should have been conserved during the evolutionary process. Hence, our goal in this section is to design the backbone of 1a2p and 2ci2 and compare the set of residues, which are conserved in our design strategy with those in the folding nucleus. As already seen in the previous section, we identify the conserved residues by monitoring the frequency with which a given residue is assigned to one of the three classes during the lowering of W controlled by suitably changing the temperature-like parameter, T, introduced in the previous section. As we said before, the tendency of one site to prefer one class over the others grows stronger as T is reduced (e.g., minimizing W). However, not all sites show this preference at the same value of T, as shown in Fig. 9, where we have shown the intensity with which protein sites in barnase are locked in the H, N, and C classes. The most conserved residues are those for which the class-locking occurs at very high temperature. It turns out that the sites involved in the locking process occupy buried positions and

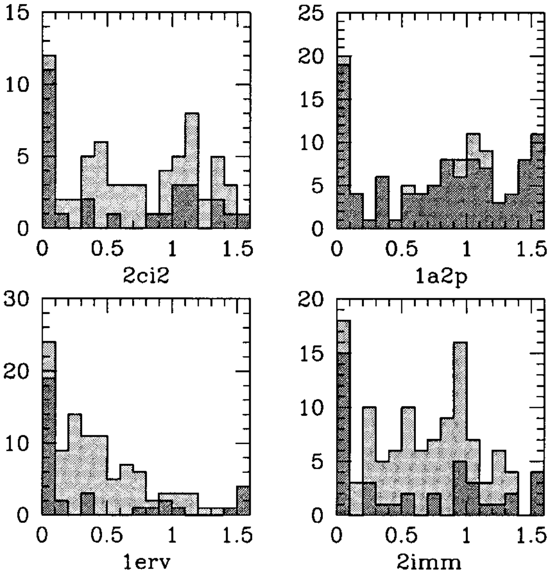

FIGURE 8 Distribution of the angles (in radians) between amino acid frequency vectors for designed sequences and aligned sequences for all the sites (*light gray*) and for conserved sites (*dark gray*). For this plot we considered conserved sites with entropy <ln(1.5).

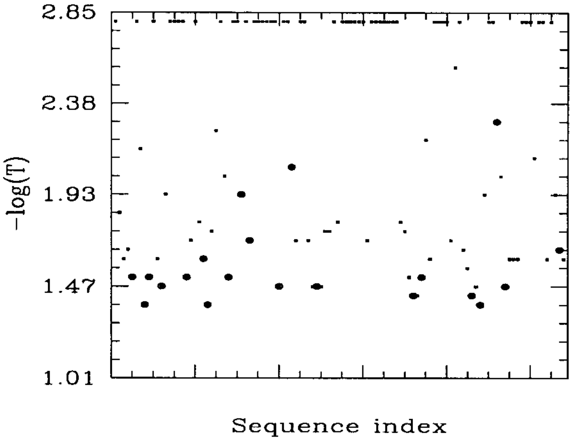

FIGURE 9 Quenched index versus sequence index for barnase. Quenched index is defined here as the first index for which the relative frequency for the hydrophobic class is >0.5. Circled dots represent sites belonging to *core1*, *core2*, or *core3* (Serrano et al., 1992).

are consistently assigned to the hydrophobic class. A visual inspection of Fig. 9 reveals that sites that are first locked in barnase correlate well with the hydrophobic *core1*, which Fersht identified as the initiator of the folding transition.

An excellent agreement with experimental findings is also observed for 2ci2, where key sites have been pinpointed through mutagenesis experiments and measurements of ϕ-values (Itzhaki et al., 1995). The key sites have been identified as those positions which are the highest rank in order of early locking. As visible in Fig. 10, the most conserved sites in our design scheme include those found to be crucial in the folding process. Again, these striking results serve a twofold purpose. On one hand, they confirm the validity of the present design approach; on the other, they also show some of its possible applications in connection with the prediction of folding nucleus.

SUMMARY

To summarize, we carried out automated protein design attempts over some PDB conformations by introducing several novel strategies to identify optimal energy-cost functions and select putative design solutions. A mere comparison of designed sequences with the PDB ones gives a success rate between 40% and 55% when working with three classes of amino acids: a value well above the random-guessing threshold. This success rate is not improving by introducing more sophisticated energy functions, suggesting that important features of real proteins are neglected by short-range Hamiltonians. Nevertheless, a statistical analysis of a wider set (nonextremal) of possible solutions shows how the design procedure could be used to correctly predict, with a high confidence, at least a subset of protein sites. These residues can be related to the conserved sites obtained by a statistical analysis of naturally occurring homologous sequences. Moreover, for two specific proteins (barnase and chymotrypsin inhibitor), these highly predictable sites correspond, with very good precision, to the folding nucleus, which is crucial for the folding process.

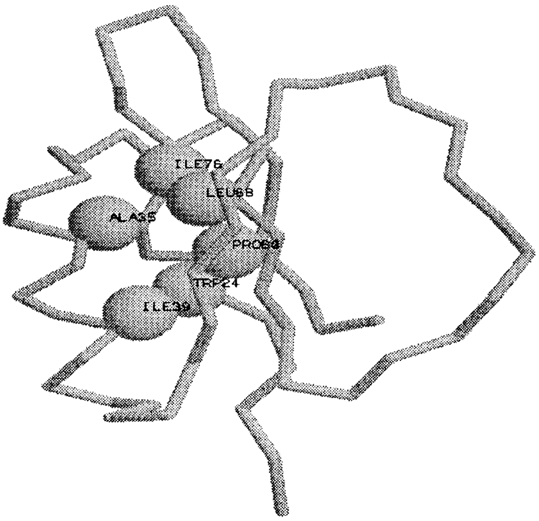

FIGURE 10 Backbone for the CI2 with the six most conserved residues in our design attempts. Three of them (Ala-35, Ile-76, Leu-68) are indicated by Itzhaki et al. (1995) as the most important in the folding process.

APPENDIX 1: DETERMINATION OF THE WEIGHT FUNCTIONS

Two-body energy

We estimated the average contact maps $\langle\Delta_{ij}^{(2)}\rangle$ and $\langle\Delta_{ijk}^{(3)}\rangle$ by considering as a set of possible competing configurations an ensemble of structures extracted from the PDB. We analyzed $N = 116$ proteins (with length ranging from 36 to 296) and for each conformation, Γ_n, we computed the corresponding value of the contact matrix $\Delta_{ij}^{(2)}(\Gamma_n)$. If the structures had the same length, $\langle\Delta_{ij}^{(2)}\rangle$ could be estimated by simple averaging:

$$\langle\Delta_{ij}^{(2)}\rangle = \frac{1}{N}\sum_{n=1}^{N}\Delta_{ij}^{(2)}(\Gamma_n)\,. \tag{11}$$

However, because we are working with proteins of different length, we can expect a dependence of $\langle\Delta_{ij}^{(2)}(\Gamma_n)\rangle$ on the length of the chains. To investigate this possibility we first notice that $\langle\Delta_{ij}^{(2)}(\Gamma_n)\rangle$ mainly depends on the sequence separation $k = |j - i|$ (at least for small k) between the amino acids along the chain more than from the position along the chain and from the length of the protein (see Fig. 11).

Let us now compute the average $\langle\Delta_k^{(2)}\rangle$ value of this contact frequencies according to

$$\langle\Delta_k^{(2)}\rangle = \frac{1}{N}\sum_{n=1}^{N}\langle\Delta_{i,j}^{(2)}(\Gamma_n)\rangle_{i-j=k}\,, \tag{12}$$

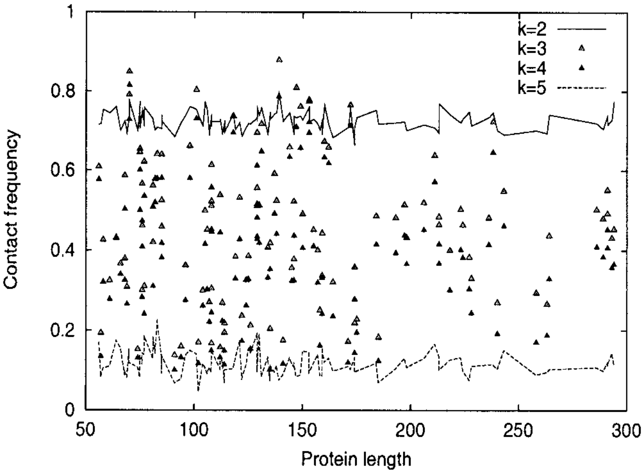

FIGURE 11 Contact frequency for different values of the amino acid separation k as a function of the length the protein. For $k = 3$ and $k = 4$ the fluctuations are large and depend on the protein family (α or β) considered α-protein or β-protein. For all the k values there is no significant dependence on the protein length.

where $\langle \cdots \rangle_{i-j=k}$ represents the arithmetic average over all the contacts with a given sequence separation k for a given protein. Then, we notice that it is a rapidly decaying function of the chemical distance, k (see Fig. 1).

We can then estimate $\langle \Delta^{(2)}_{ij} \rangle$ according to the rules:

$$\langle \Delta^{(2)}_{ij} \rangle = \begin{cases} \langle \Delta^{(2)}_{k} \rangle & k < k_0 \\ \Delta^{(2)}_{0} & k \geq k_0. \end{cases} \tag{13}$$

where k_0 is a cutoff distance that we fixed equal to 16. The value $\langle \Delta^{(2)}_{k} \rangle$ can be estimated numerically from the data bank through Eq. 12, whereas $\Delta^{(2)}_{0}$ should be determined according to the length of the chain.

Indeed, the dependence of the total number of contacts, $\Sigma_{i<j}\Delta^{(2)}_{ij}(\Gamma_n)$, is well approximated by a linear function of the length, or number of amino acids, L_n, of Γ_n. Thus, using this linear dependence on L_n and Eq. 13 we are able to determine $\Delta^{(2)}_{0}$.

Three-body energy

The average contact map $\langle \Delta^{(3)}_{ijk} \rangle$ can be determined in an analogous way. For a conformation Γ we define the total number of three-body contacts as

$$N^{3}_{c}(\Gamma) = \sum_{i<j<k} \Delta^{(3)}_{ijk}(\Gamma)\,. \tag{14}$$

Similarly to the former case, this number of contacts can be fitted by a linear relation. In this larger parameter space $\langle \Delta^{(3)}_{ijk} \rangle$ will depend on two indexes, $k1 = |j - i|$ and $k2 = |k - j|$:

$$\langle \Delta^{(3)}_{ijk} \rangle = \Delta^{(3)}(k1, k2)\,. \tag{15}$$

For $k1$, $k2 < k_0$ (that we choose on the basis of the statistical analysis to be $k_0 = 6$)

$$\Delta^{(3)}(k1, k2) = \frac{1}{N} \sum_{n=1}^{N} \langle \Delta^{(3)}_{ijk}(\Gamma_n) \rangle_{j-k=k_2, i-j=k_1} \tag{16}$$

while for $k_1 \geq k_0$ or $k_2 \geq k_0$ we assume a constant value. Here, $\langle \cdots \rangle_{j-k=k_2, i-j=k_1}$ represents the arithmetic average over all the contacts with given sequence separation k_1, k_2.

The average contact map for a generic protein will be

$$\langle \Delta^{(3)}_{ijk} \rangle = \begin{cases} \Delta^{(3)}(k1, k2) & k1, k2 < k_0 \\ \Delta^{(3)}_{0} & \text{otherwise.} \end{cases} \tag{17}$$

Using, again, that $\Sigma_{i<j<k}\langle \Delta^{(3)}_{ijk} \rangle$ is well interpolated by a linear function of L_n, we can determine $\Delta^{(3)}_{0}$ in Eq. 17 after $\Delta^{(3)}$ (k_1, k_2) for k_1, $k_2 < k_0$ have been evaluated.

APPENDIX 2: PERCEPTRON LEARNING OF THE OPTIMAL POTENTIALS

A convenient way to find the optimal potentials that satisfy inequality constraints such as those of Eq. 7 is the use of the perceptron algorithm for the optimization of a set of linear inequalities (Krauth and Mezard, 1987; van Mourik et al., 1999).

For instance, in the case of the two-body Hamiltonian, Eq. 7 can be written, using the result of Eq. 8, as:

$$\sum_{i>j=1}^{L} (\langle \Delta^{(2)}_{ij} \rangle - \Delta^{(2)}_{ij}(\Gamma)) B_2(s_i, s_j) > 0 \tag{18}$$

where L is the length of the protein. If $n_{kl}(\Gamma)$ denotes the number of contacts in the conformation Γ involving amino acids of types k and l, and $\langle n^{(2)}_{kl} \rangle$ the corresponding average computed on the set of competing configurations by using Eq. 13, Eq. 18 can be rewritten as:

$$\sum_{k>l=1}^{3} (\langle n^{(2)}_{kl} \rangle - n_{kl}(\Gamma)) B_2(k, l) = \sum_{k>l=1}^{3} a_{kl}(\Gamma) B_2(k, l) = \mathcal{F}_{\Gamma}(\vec{B}) \tag{19}$$

where the vector $\vec{B}_2$ is defined as:

$$\vec{B} \equiv (B(1, 1), B(1, 2), B(1, C), B(2, 2), B(2, 3), B(3, 3)) \tag{20}$$

Given the native state Γ and the sequence s, the six entries of a_{kl} depend only on the average properties of the decoy structures.

For a given set of M inequalities to be satisfied simultaneously, it is convenient to identify the one (related to the conformation Γ_s) that, with a given set of trial potentials, is the least satisfied one, e.g.:

$$\mathcal{F}_{\Gamma_s}(\vec{B}) < \mathcal{F}_{k}(\vec{B}) \quad k = 1, \ldots, M \quad k \neq s \tag{21}$$

Once Γ_s has been determined, one updates the trial potentials adding a quantity proportional to $a_{kl}(\Gamma_s)$, where the proportionality constant is chosen to be much smaller than one. With this new choice of the potentials, each inequality is re-evaluated and the updating cycle is repeated until $\mathcal{F}_{\Gamma_s}(\vec{B})$ (stability) reaches the maximum possible value. One is allowed to fix the scale of B values by requiring $|\vec{B}| = 1$, where the $|\cdot|$ is the usual Euclidean norm. This method can be shown to converge to an optimal solution, $\mathcal{F}^*$, which can be of either sign. If it is negative, it means that no set of potentials can be found that consistently satisfied all inequalities in the set. Otherwise, the problem is learnable and the optimal potentials are identified with those giving the highest stability.

We have extracted potentials by using the perceptron scheme with M = 31 globular proteins. The related set of inequalities has turned out to be learnable in all cases, with two- or three-body energy terms.

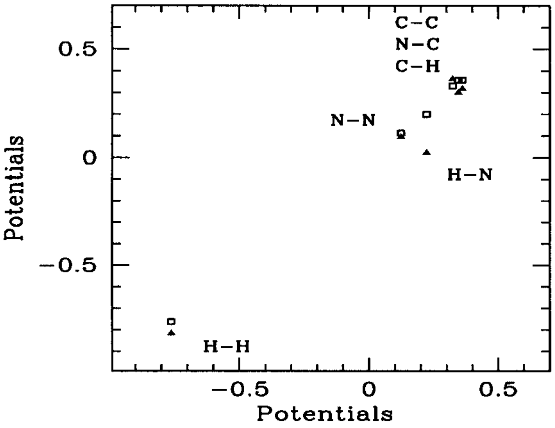

FIGURE 12 The potentials $\vec{B}$ determined using a set of 15 proteins and another set of 16 proteins (see Table 2) are plotted versus the same potentials determined by the whole set of 31 proteins. The correlation between the potentials obtained with the two sets and the largest one is nearly perfect (ideally, points should lie on the diagonal). Using the whole set of Table 2 we found $\vec{B}$ = (0.12, 0.22, 0.36, −0.76, 0.35, 0.32). Potentials are here sorted as in Eq. 20, where 1, 2, 3 refer respectively to classes P, H, C.

For the two-body energy we have extracted a first set of potentials using the 15 proteins and a second one with the remaining 16. The two sets of potentials are plotted one versus the other in Fig. 12, showing a extremely good correlations.

This validates the conclusion that an interaction matrix *B* depending only on six parameters can be determined with a dozen nonredundant globular proteins. Similar results have been obtained with the three-body energy (Serrano et al., 1992).

We thank Marco Punta for many useful discussions. This work was supported by INFM (PAIS project) and MURST (COFIN99).

REFERENCES

Anfinsen, C. 1973. Principles that govern the folding of protein chains. *Science.* 181:223–239.

Cecconi, F., C. Micheletti, P. Carloni, and A. Maritan. 2000. The structural basis of antiviral drug resistance. *Proteins Struct. Funct. Genet.* in press.

Chan, H., and K. A. Dill. 1993. The protein folding problem. *Physics Today.* 46:24–32.

Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequences and structures in proteins. *EMBO J.* 5:823–826.

Crippen, G. 1991. Prediction of protein folding from amino acid sequence over discrete conformation space. *Biochemistry.* 30:4232–4237.

Dahiyat, B., and S. Mayo. 1997. De novo protein design: fully automated sequence selection. *Science.* 278:82–87.

Deutsch, J., and T. Kurosky. 1996. New algorithm for protein design. *Phys. Rev. Lett.* 76:323–326.

Dill, K. A., S. Bromberg, K. Yue, K. Fiebig, D. Yee, and P. Thomas. 1995. Principles of protein folding: a perspective from simple exact models. *Protein Sci.* 4:561–602.

Dima, R., G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan. 2000. Extraction of interaction potentials between amino acids from native protein structures. *J. Chem. Phys.* 112:9151–9166.

Fersht, A. 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman, New York.

Fersht, A. R. 1995. Optimization of rates of protein folding: the nucleation condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA.* 92:10869–10873.

Hoang, T., and M. Cieplak. 2000. Molecular dynamics of folding of secondary structures in go-like models of proteins. *J. Chem. Phys.* 112:6851–6862.

Huang, E., P. Koehl, M. Levitt, R. Pappu, and J. Ponder. 1998. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins: Struct; Funct; Genet.* 33:204–207.

Itzhaki, L. S., D. E. Otzen, and A. R. Fersht. 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* 254:260–288.

Kamtekar, S., J. Schiffer, H. Xiong, J. Babik, and M. Hecht. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science.* 262:1680–1685.

Krauth, W., and M. Mezard. 1987. Learning algorithms with optimal stability in neural networks. *J. Phys. A.* 20:L745–L752.

Lau, K. F., and K. A. Dill. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules.* 22:3986–3997.

Maiorov, V. N., and G. M. Crippen. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888.

Maritan, A., C. Micheletti, and J. Banavar. 2000a. Role of secondary motifs in fast folding polymers: a dynamical variational principle. *Phys. Rev. Lett.* 84:3009–3012.

Maritan, A., C. Micheletti, A. Trovato, and J. Banavar. 2000b. Optimal shapes of compact strings. *Nature.* 406:287–290.

Micheletti, C., J. Banavar, A. Maritan, and F. Seno. 1999a. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* 82:3372–3375.

Micheletti, C., A. Maritan, and J. R. Banavar. 1999b. A comparative study of existing and new design techniques for protein models. *J. Chem. Phys.* 110:9730–9738.

Micheletti, C., F. Seno, A. Maritan, and J. Banavar. 1998a. Design of proteins with hydrophobic and polar amino acids. *Proteins: Struct; Funct; Genet.* 32:80.

Micheletti, C., F. Seno, A. Maritan, and J. Banavar. 1998b. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys. Rev. Lett.* 80:2237.

Micheletti, C., F. Seno, A. Maritan, and J. R. Banavar. 1999c. Strategies for protein folding and design. *Ann. Combinatorics.* 3:439–458.

Morrisey, M., and E. Shakhnovich. 1996. Design of proteins with selected thermal properties. *Folding and Design.* 1:391–405.

Pabo, C. 1983. Designing proteins and peptides. *Nature.* 301:200.

Park, B., and M. Levitt. 1996. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258: 367–392.

Quinn, T. B., N. B. Tweedy, R. W. Williams, J. S. Richardson, and D. C. Richardson. 1994. De novo design, synthesis and characterization of a beta sandwich protein. *Proc. Natl. Acad. Sci. USA.* 91:8747–8751.

Rossi, A., C. Micheletti, and A. Maritan. 2000. A novel iterative strategy for protein design. *J. Chem. Phys.* 112:2050–2055.

Sander, C., and R. Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins.* 9:56–68.

Seno, F., A. Maritan, and J. Banavar. 1998a. Interaction potentials for protein folding. *Proteins: Struct; Funct; Genet.* 30:224–248.

Seno, F., C. Micheletti, A. Maritan, and J. Banavar. 1998b. Variational approach to protein design and extraction of interaction potentials. *Phys. Rev. Lett.* 81:2172.

Seno, F., M. Vendruscolo, J. Banavar, and A. Maritan. 1996. Optimal protein design procedure. *Phys. Rev. Lett.* 77:1901–1904.

Serrano, L., J. T. Kellis, P. Cann, A. Matousheck, and A. R. Fersht. 1992. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* 224:783–804.

Shakhnovich, E. I. 1994. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* 72:3907–3910.

Shakhnovich, E., V. Abkevich, and O. Ptitsyn. 1996. Conserved residues and the mechanism of protein folding. *Nature.* 379:96–98.

Shakhnovich, E., and A. Gutin. 1993. A new approach to the design of stable proteins. *Protein Eng.* 6:793–800.

Street, A. G., and L. S. Mayo. 1999. Computational protein design. *Structure with Folding and Design.* 7:R105–R109.

Sun, S., R. Brem, R. Chan, and K. Dill. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.* 8:1205–1213.

van Mourik, J., C. Clementi, A. Maritan, F. Seno, and J. Banavar. 1999. Determination of interaction potentials of amino acids from native protein structures: tests on simple lattice models. *J. Chem. Phys.* 110: 10123–10133.

Vendruscolo, M., and E. Domany. 1999. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* 109:11101–11108.

West, M. W., W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht. 1999. De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA.* 96:11211–11216.

Zou, J., and J. G. Saven. 2000. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. Mol. Biol.* 296:281–294.