Protein design: a perspective from simple tractable models

Eugene I Shakhnovich

Recent progress in computational approaches to protein design builds on advances in statistical mechanical protein folding theory. Here, the number of sequences folding into a given conformation is evaluated and a simple condition for a protein model's designability is outlined.

Address: Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138, USA.
E-mail: eugene@belok.harvard.edu

Folding & Design 01 June 1998, **3**:R45–R58
http://biomednet.com/elecref/13590278003R0045

The protein folding problem has two components: the 'direct' folding problem (i.e. folding) and the 'inverse' problem (i.e. protein design). The main issue of the direct protein folding problem is to understand the basic physical chemistry of how protein sequences determine their structure. The long-range goal of these studies is to predict protein conformations from a given sequence. The direct protein folding problem has received much attention recently and considerable progress has been made towards understanding the general principles that govern the folding of protein chains [1–4]. Using the language of bioinformatics, one can define the folding problem as mapping the space of sequences onto the space of structures. The inverse protein folding problem involves finding a sequence that folds into and is stable in a given conformation at a given temperature (Figure 1). Again using the language of bioinformatics we can say that this corresponds to mapping the space of structures onto the space of sequences.

It is clear that the two problems are closely related to each other: a better understanding of the principles of protein folding makes it possible to clarify which features of protein sequences are necessary (as well as sufficient) for stability and fast folding; in other words, the features that make a protein a protein. Such understanding focuses the attention of designers on emphasizing these crucial features of folding sequences.

The direct folding problem

The experimental approaches to protein structure determination have been very successful, providing a wealth of structural information. Although the growing flow of genomic information makes the development of theoretical approaches to predict protein conformation even more desirable, there is an experimental 'shortcut' using X-ray crystallography or NMR that can be taken to reach the solution of the direct protein folding problem.

The protein design problem

The situation with the inverse problem is very different. Most of the present experimental approaches enjoyed only limited success, providing polypeptides that in most cases fold into compact but mostly disordered conformations of molten-globule-like species (see e.g. [5]). It is quite possible that limitations in experimental design result from a relatively low synergism between experiment and theory. An important success story based on such synergism of theory and experiment is given in [6]; in this case, theoretical analysis has helped to guide a design effort that resulted in a small protein that folded into a predicted 'target' conformation. This work clearly demonstrates the importance of theory in protein design. A limitation of the approach reported in [6] is that it requires complete enumeration of sequence candidates — a problem that explodes exponentially with chain length and therefore limits this valuable approach to relatively short chain lengths. The successes and limitations of the work of Dahiyat and Mayo [6] call for further refinement of theoretical approaches to protein design, some of which will be outlined in this review.

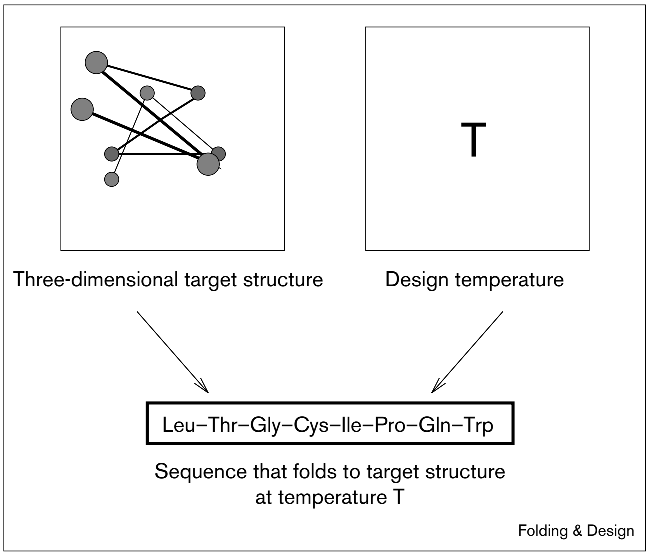
It is important to note that the bottleneck in protein design is not on the synthetic side, but rather in the fundamental problem that researchers generally do not know which sequences to synthesize. Because the number of possible sequences is enormous and the fraction able to fold into protein-like structures is negligible (see below), the probability that one will 'hit' a correct sequence by chance is vanishingly low. Of course, there exist clever experimental approaches, such as phage display [7], that bias experimental sequence searches towards better candidates. In our view, however, convincing success in protein design will come with reliable theoretical approaches that will make it possible to find sequences that fold uniquely into a desired conformation. Perhaps this goal alone justifies all the effort that has been put into protein folding theory over the past few years.

In this review, I discuss how recent advances towards understanding protein folding can help us to design protein sequences and to understand their natural evolution.

Mapping structures into sequences: how many protein sequences are there?

The computational approach to protein design aims to find sequences that fold to a given structure in a particular model. The fundamental question is whether or not there is any solution to this problem (for a model, of course, because we know that there is one for proteins) and if there is, how many solutions are there? In other words, how many sequences can fold into a given conformation? This

Figure 1


T
Three-dimensional target structure
Design temperature
Leu–Thr–Gly–Cys–Ile–Pro–Gln–Trp
Sequence that folds to target structure
at temperature T
Folding & Design

A schematic representation of the 'inverse' protein folding problem (i.e. protein design; taken from [52]). Given the target three-dimensional structure and the selected temperature a sequence is found that folds at this temperature into the given conformation and is stable in this conformation.

question can be addressed only if we understand the features that a folding sequence should have. Such understanding builds on recent developments in protein folding theory, which elucidated some of the properties of folding sequences [8–11].

The thermodynamic requirements of folding

According to a thermodynamic hypothesis [12] sequences that fold into a given structure have the lowest energy (potential of mean force) in that structure, compared with energies of decoys (i.e other conformations for the same sequence). The 'consistency principle' [13] and the 'principle of minimal frustrations' (PMF) by Bryngelson *et al.* [2] apparently posited that the necessary condition for protein stability and fast folding is that the native state has an energy that is much lower than the energies of the bulk of misfolded states (decoys). In modern language, one can say that the PMF is actually equivalent to the requirement of a large energy gap in protein-like models.

The results of the analytical microscopic theory of heteropolymer folding [14–17] as well as numerical studies [9,10, 18] in lattice models are consistent with the PMF. More specifically, it was shown that in order for a sequence to fold into a given native structure, its energy in that structure should fall below a certain threshold E_c (the energy at which the density of states for decoys vanishes): at $E \geq E_c$ the density of states is very high so many decoys belong to that energy range (Figure 2). The probability that there will be a decoy, structurally unrelated to the native conformation and with an energy $E < E_c$, has been estimated in the Appendix to [10] to be $\exp[(E - E_c)/T_c]$, where T_c is the temperature of the thermodynamic freezing transition in a random heteropolymer. (The thermodynamic freezing transition is defined as the temperature at which the coarse-grained entropy of a polymer vanishes [14,19].) Thus, if a sequence folds into a given structure with energy E, the probability that there will be a structurally dissimilar decoy having an equal or lower energy falls off exponentially and for sequences that fold into the target structure with sufficiently low energy E, such that $E_c - E >> T_c$, the target structure will almost certainly be a unique ground state conformation. Further studies showed that the pronounced 'stability gap' $E - E_c$ is also sufficient to provide fast folding for lattice model proteins of considerable length (more than 100 monomers; [18,20]), consistent with the PMF [21].

Thus, a possible search criterion for folding sequences is a large (many kT_c) stability gap. With this, the issue of how many sequences are able to fold into a given conformation (the degeneracy of the protein code) is reduced to the question of how many sequences $N(E)$ exist that have the energy $E < E_c$ in a given structure:

$$N(E) = \sum_{\text{seq}} \delta(H_{(\text{seq,conf})} - E) \qquad (1)$$

where $H_{(\text{seq,conf})}$ is the energy of a particular sequence in the target conformation. δ means that the summation is taken over all sequences that have energy E in the native conformation.

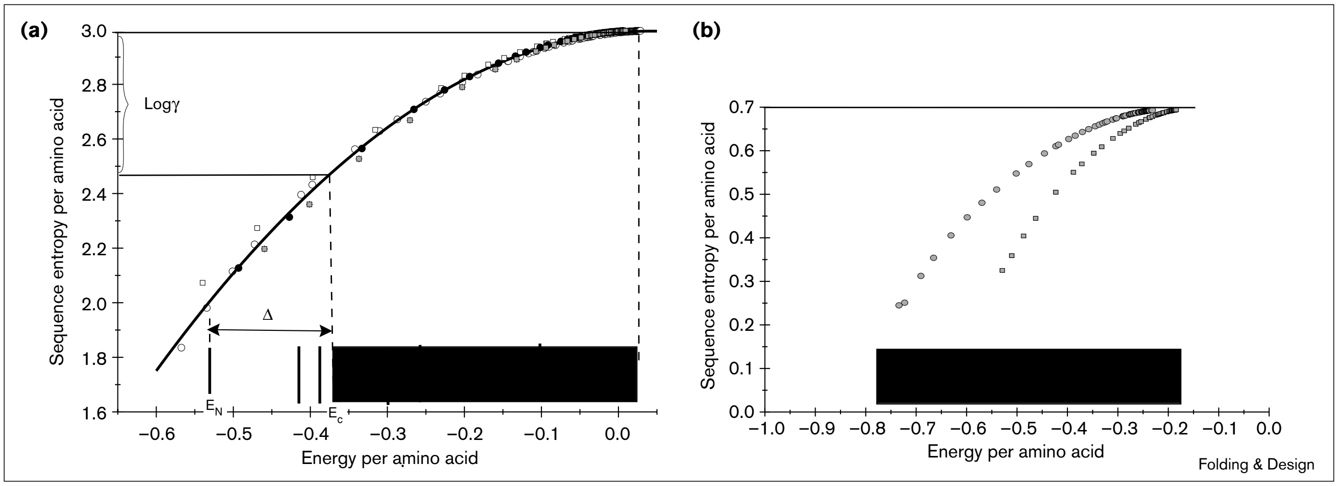
An example of the stability gap condition

In a particular example that has received much attention in the past [1,22–25], H is a contact potential:

$$H(\{\sigma\},\{r\}) = \sum_{i<j}^{N} [U(\sigma_i,\sigma_j)]\Delta(r_i,r_j) \qquad (2)$$

where N is the number of residues in the chain and σ_i characterizes the type of monomer i so that the sequence of monomers is defined as a sequence of symbols $\{\sigma\}$. There are 20 types of amino acid so $\sigma_i = 1...20$. The parameters $U(\sigma_i,\sigma_j)$ determine the magnitude of the contact interaction between monomers of type σ_i and σ_j; several sets of such parameters have been published [22,23, 26,27]. A simple approximation of the conformation of a chain uses residue representation: a residue i is assigned a one-point location variable r_i (it can be a geometrical center of the sidechain or a coordinate of its Cα atom or Cβ atom). $\Delta(r_i,r_j) = 1$ if residues i and j are in contact and $\Delta(r_i,r_j) = 0$ otherwise. For protein structures, a reasonable definition of a contact is when the distance between their Cα/Cβ atoms is < 6.5 Å [22]. For lattice model proteins, the definition of a contact is even simpler: two amino acids that are nearest neighbors on the lattice but are not sequence neighbors are considered contacting.

Figure 2



Folding & Design

Degeneracy of the protein code. **(a)** The designable model, for which $m_{eff} > \gamma$. Many sequences ($\sim e^{1.9N}$ in the present example) exist that have a low energy E_N in the target conformation with a pronounced stability gap $\Delta = E_N - E_c$. Such sequences are expected to fold fast to the native conformation. Data points correspond to the direct calculation of sequence entropy from Monte-Carlo simulations in a range of selective temperatures (keeping the amino acid composition the same as in the native sequence). Different symbols correspond to different proteins. **(b)** The non-designable model, where $m_{eff} < \gamma$. No sequences that fold uniquely to the ground state can be found. The model runs out of sequences at energies that are not low enough to ensure a large gap between the native structure and misfolded decoys. The data points represent the entropy of Monte-Carlo design simulations for two proteins' HP models [41]: upper curve, myoglobin (PDB code 1mbn); and lower curve, plastocyanine (PDB code 1pcy). Amino acids were categorized into 'H' and 'P' classes as explained in [41]. The more pronounced difference between proteins results from the difference in their average hydrophobicities (i.e. the fraction of hydrophobic residues in their sequences). For both (a) and (b) the entropy and energy are shown normalized per amino acid residue: $s_{seq} = S_{seq}/N$ and $e_N = E_N/N$, respectively. The horizontal insert is given to show schematically the generic representation of the density of states in conformational space, as predicted by the heteropolymer theory [14,19]. The range of energies at which the density of non-native decoys is high is shown; in (a), a few low-energy conformations (shown as discrete lines in the insert) that lie below the boundary of the continuous spectrum E_c (the energy at which the density of states for decoys vanishes) represent lowest energy decoys. The solid line is the analytical formula of Equation 3. The average energy E_{av} and the dispersion D were calculated as explained in the text using the Myazawa–Jernigan set of parameters (Table 4 in [22]). Simulations using another parameter set [23] gave identical results. The average energy of sequences in the target structure E(T) was evaluated from long simulation runs. Equation 11 was applied to obtain sequence space entropy.

N(E) in Equation 1 can be evaluated using the technique that represents the Dirac δ function in Equation 1 via Fourier transform, which expands appearing exponentials up to the second order, sums over all sequences and re-exponentiates the result. The final result of the calculation can be expressed in terms of the 'entropy' in sequence space:

$$S_{seq}(e) = \ln N(E) = \log(m_{eff}) - \frac{(E - E_{av})^2}{2ND^2} \qquad (3)$$

where m_{eff} is the effective number of types of amino acids:

$$m_{eff} = \exp(-\sum_{i=1}^{20} p_i \ln p_i) \qquad (4)$$

For example, if all types of amino acids are equally represented so that $p_i = 1/20$ for any i then $m_{eff} = 20$. In the opposite case when, for example, $p_1 = 1$ and $p_i = 0$ for any $i = 2...20$, $m_{eff} = 1$, which makes clear sense because this situation corresponds to a homopolymer. E_{av} is the average (over all conformations) energy of interactions per amino acid and D is the dispersion of interaction energies per contact. E_{av} is calculated as an average interaction energy over all possible contacts; E_{av} depends on the amino acid composition, but not on the details of the sequence. D is the dispersion of contact energies and is also calculated over all possible contacts.

Calculation of these quantities does not require simulations or enumerations in conformational space. Certain geometrical properties, which may restrict the types of possible contacts, should, however, be taken into account. For example, for a cubic lattice an important property is that contacts are only possible between units with opposite parity. This 'even–odd' rule should be taken into account in estimating E_{av} and D for the cubic lattice model.

The question of how many sequences fold into a given structure was first addressed by Finkelstein *et al.* [28] who postulated the distribution given in Equation 3. According to the heteropolymer theory [14,19,21,29], the density of states of a three-dimensional heteropolymer (the number of conformations having energy in a given range) follows the random energy model distribution:

$$W(E) = \gamma^N \exp -\left(\frac{(E - E_{av})^2}{2ND^2} \right) \qquad (5)$$

where γ is the number of conformations per monomer. The energy at which the chain runs out of states (the boundary of the continuous spectrum E_c in the insert in Figure 2) is estimated from the condition W(E) ~1, such that:

$$E_c - E_{av} = N(2\ln\gamma)^{1/2}D \tag{6}$$

As explained above, a necessary condition that determines a folding sequence is that its energy in the native state is $E < E_c$. Such sequences should exist; in other words, $S_{seq}(E < E_c) > 0$. It follows from Equations 3 and 6 that this condition can be satisfied only when:

$$m_{eff} > \gamma \tag{7}$$

Apparently, there is another threshold energy, E_{lowest}, such that there are no sequences that have an energy in the native state lower than E_{lowest}. A possible crude estimate of E_{lowest} can be obtained from the condition that at this energy the system runs out of sequences. Mathematically, this is equivalent to the condition $S_{seq}(E_{lowest}) = 0$. It is quite possible, however, that this is an overestimate and the actual boundary of lowest possible energies in a sequence model may be higher than estimated from the entropy condition below.

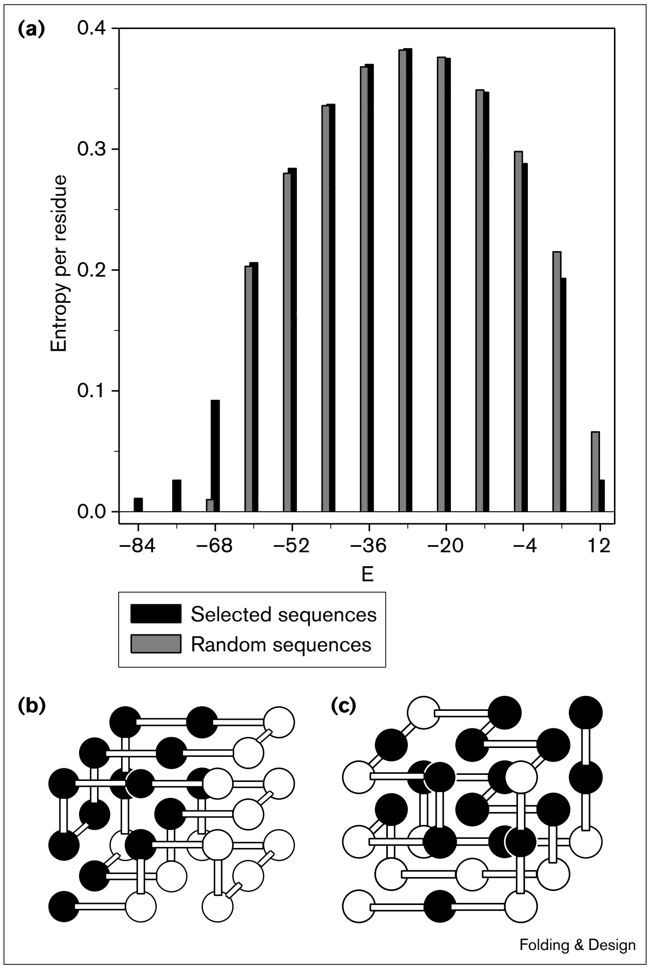
Thus, the upper bound estimate of the maximal possible gap $E_{lowest} - E_c$ is:

$$G_{max} = N \ln \frac{m_{eff}}{\gamma}(2D^2)^{1/2} \tag{8}$$

Designed versus random sequences

A specific simple example to clarify the main concepts of this analysis is presented in Figure 3. It shows the energy spectrum or densities of states (the logarithm of the number of conformations having a given energy) for the designed sequence and a random sequence having the same composition. Comparing this spectrum with the one presented schematically in the insert in Figure 2 one should keep in mind that for the model that has only two types of amino acid the spectrum is apparently discrete because possible values of energy are determined by numbers of contacts of different types, which are obviously integer numbers (a straightforward generalization of heteropolymer results to this discrete case is given in [30]). The occupancy at each energy level (i.e. how many conformations have that energy) is, however, different for different levels. Specifically, there may be energy levels that are highly populated (i.e. a multitude of conformations have that energy). There are also empty low-energy levels, which can be filled only for special sequences (i.e. only certain sequences can have such an ‘unusually’ low energy in their native conformations). The designed sequence shown in Figure 3 has the lowest possible energy for the model ($E_N = E_{lowest} = -84$) in its unique native conformation.

Figure 3



Folding & Design

An example to clarify the concept of the density of states analysis. **(a)** The density of states (energy spectrum) for the ensemble of fully compact conformations of the 27-mer model for a best-designed sequence (black bars) and a random sequence (gray bars). Each bar corresponds to the entropy per residue – the logarithm of the number of all conformations having a given energy divided by the number of residues (27 in this case). The density of states plots are derived from exhaustive enumeration of all 103,346 compact conformations of the 27-mer [29]. For simplicity, only two types of monomers are used (‘black’ and ‘white’) with nearest neighbor ‘color specific’ interactions: $E_{BB} = E_{WW} = -3$; $E_{BW} = -1$ [9,25]. The best-designed and random sequences have the same composition (13B, 14W). Although this interaction matrix may be not quite realistic for real proteins, it is useful for clarifying basic concepts presented in this review. Obviously the lowest energy conformation is the one that maximizes the number of favorable ‘same color’ (SC) contacts. **(b)** The target structure and the sequence that has the minimal possible energy ($E_{lowest} = -84$; all 28 contacts are SC) in that structure. This structure presents a unique ground state for the designed sequence. The black bar in (a) for the designed sequence corresponding to the energy $E_N = -84$ is slightly exaggerated to make it visible. **(c)** The ground state for a quasi random sequence.

It can be seen clearly in Figure 3 that the spectra for the random and the designed sequences differ only at the low energy part: at energies that are ≥ –60 both the random sequence and the designed sequence have almost identical spectra (i.e. this part of the spectrum is sequence independent). Quantities that are sequence independent are called self averaging [2,29,31]. According to the heteropolymer theory [14,19,21,29] the density of states is self-averaging at energies E_c and higher while the low-energy part at $E < E_c$ is sequence specific. The low-energy non-self-averaging part of the spectrum represents an energetic fingerprint of a sequence.

It follows that for this model $E_c = -60$. Note also the concave shape at the left wing of the spectrum for the designed sequence; it is a signature of a cooperative transition [13]. The cooperativity of a transition (e.g. its widths) is directly related to the value of the relative gap $g = (E_N - E_c)/E_N$. For this model, $m_{eff} \sim 2$. Only compact conformations are considered, so $\gamma = 103346^{1/26} \sim 1.7$. The relative gap is $g = -0.33$.

Lessons for design

The statistical-mechanical analysis suggests a number of lessons.

Lesson 1: the design problem may be easier than the folding problem

In a protein-like model where $m_{eff} > \gamma$ there is a large (exponential in the chain length N) number of sequences that have a sufficiently large energy gap $G \sim ND$ to fold reliably into the target structure. Unlike folding, in which a unique ground state solution is sought, in design any sequence having a sufficient (not necessarily the greatest possible) energy gap [8,9] folds cooperatively into the target conformation if the temperature is not too low [32]. Although the number of folding sequences is large, the fraction of folding sequences (i.e. the probability to pick up a cooperatively folding sequence from the ensemble of random sequences) is quite low. This makes the design problem nontrivial.

Lesson 2: the number of amino acid types may be important when determining the designability of a protein model

The models in which the number of types of amino acid (m_{eff}) are small are ‘undesignable’. This means that even the best sequences designed for these models have an energy in the native state higher than E_c (i.e. decoys with an energy lower or equal to the energy of the designed sequences in the native state are present in such models). Apparently no folding is possible in this case because the native structure is not unique. An example of such an undesignable model is the so-called HP model [33].

Lesson 3: ‘stiffer’ chains provide greater energy gaps and are therefore more designable

The fundamental relationship for a designable model, the condition presented in Equation 7 can be enforced either by increasing the number of amino acid types or by decreasing γ (i.e. by decreasing the number of conformations per monomer). There are a number of ways to decrease γ: formation of secondary structure; forcing the conformational ensemble of a chain to the set of compact conformations (by introducing additional non-specific attraction, Figure 3); and biasing the conformations to carry certain structural features (as in threading). The example given in Figure 3 shows that even the ‘two-letter’ model may sometimes have a non-degenerate native state (but a very small gap) if its configurational space is restricted to only compact conformations. When a full ensemble is considered, the ground state of HP sequences become multiple degenerate [9,33,34]. Apparently the number of all conformations (per monomer) γ_{all} is greater than the number of compact conformations $\gamma_{compact}$ so that the condition in Equation 7 is violated for the HP model when all conformations are considered. On the other hand, the two-letter models that are restricted to maximally compact conformations are only just ‘on the borderline’ of the validity of the condition in Equation 7.

Lesson 4: ‘designing out’ decoys may be necessary for two-dimensional models

The key to successful protein design is to find sequences that have a low energy for the native state without optimizing decoys at the same time. This factor increases the energy gap or, equivalently, increases the thermal probability to be in the native state (see below). To this end the ‘ruggedness’ of the conformational space of three-dimensional random heteropolymers (as exemplified by the equivalence between heteropolymers and the random energy model, REM; [14,15,21]) is important. According to the REM, most low-energy decoys are structurally different from the native state (except the ones that represent small fluctuations around the native conformation — the native state ensemble). To this end, optimization of the native conformation energy (i.e. making the native contacts stronger) does not affect the low-energy, structurally dissimilar decoys (see Figure 3). This makes designing ‘in’ on the background of decoys that are unaffected by sequence selection an efficient way to increase the energy gap. We should emphasize that this is true only for three-dimensional models; in two dimensions the optimization of the native states gives rise to optimization of numerous partly folded low-energy decoys, making the native state unstable (in contrast to the three-dimensional case in which partly folded decoys have a high energy). The physical reason for such dramatic dependence on space dimensionality is given in [35,36]; particularly see the Appendix to [36]. In three-dimensional compact chains non-local contacts dominate, whereas in two-dimensional chains local contacts dominate. It was pointed out by several authors [37–39] that some special three-dimensional target conformations (crumpled globules [40]) may be as ‘undesignable’ by simple methods as two-dimensional models, for the same reason — prevalence of local contacts.

Stochastic optimization in sequence space: a simple model solution for the design problem

The major lesson from the statistical mechanical theory is that many solutions of the design problem exist. A crucial question of practical importance is how to find such solutions. To this end, a number of approaches (reviewed in this section) of various complexity and efficiency have been suggested.

Increasing the thermal probability for the native state

It is clear that all that is needed for successful design is to find a sequence $\{\sigma_i\}$ that has a high thermal probability to be in the native state:

$$P(T) = \frac{e^{-\frac{H(\{\sigma_i\},\{r_i^0\})}{k_b T}}}{Z(\{\sigma_i\})} \qquad (9)$$

where the native state is characterized by the set of coordinates of its residues $\{r_i^0\}$, H is the energy of a given sequence in a given conformation (cf. Equation 2), Z is a partition function of the chain:

$$Z(\{\sigma_i\}) = \sum_{r_i} e^{-\frac{H(\{\sigma_i\},\{r_i^0\})}{k_b T}} \qquad (10)$$

where the summation is taken over all conformations of the chain $\{r_i\}$, T is temperature, and k_b is the Boltzmann constant.

As presented by Equations 9 and 10, the problem of design is of great complexity because it involves a search in both conformational and sequence spaces. (The search in conformational space is needed to determine the partition function.) In other words, the 'exact' solution of the design problem that includes exhaustive searches in conformational and sequence spaces would require $(m_{eff}\gamma)^N$ 'trials' — a prohibitive number for any model of practical interest.

This calls for the development of approximations that would allow one to avoid an exhaustive search both in sequence space and in conformational space. The simplest approach of this kind was proposed in 1993 in [9]. It is based on the following ideas.

The optimization of stability is equivalent, in a simplest case, to the maximization of the energy gap g defined above (see Figure 1 of [9] for a qualitative explanation of this fact). The boundary of the continuous spectrum E_c is a self-averaging quantity, i.e. it depends on amino acid composition only while the lower part of the spectrum $E < E_c$ is highly sequence specific. This conjecture from heteropolymer statistical mechanics was shown to be correct for simple exact models, such as the one shown in Figure 3. It follows that the desired design results can be obtained by selection of sequences that have a low energy in the target conformation at a given amino acid composition. It is clear that this statement is equivalent to the assumption that the partition function Z (more precisely, the contribution to Z from non-native-like decoys) in Equation 9 depends primarily on amino acid composition rather than on sequence. The analysis using the REM approximation suggests that this conjecture is valid at high enough temperatures $T > T_c$, where T_c is the temperature of the 'freezing' [14,19,21] transition in a random heteropolymer having the same amino acid composition. A lucid discussion of this point and further details can be found in [19].

The gap optimization in sequence space can be achieved by any stochastic algorithm. In the case of sequence design, the energy landscape in sequence space is 'smooth' [9,41] so there is no complicated search problem. Thus, a simple Monte-Carlo algorithm would suffice [8,18,20,42].

An *experimentum crucis* to test the statistical-mechanical approach to sequence design is to pick an arbitrary conformation and design a sequence that is expected to fold into that conformation. A proof of concept for a design method is an actual folding simulation of a designed sequence, starting from an arbitrary random-coil conformation. If the designed sequence converges to the target conformation and never encounters grossly misfolded conformations with an energy lower than the target conformation, then they may be stable in the target state and the design is successful.

This program has been carried out in [9,18], in which random mutations preserving the amino acid composition (monomer swaps) were introduced under Metropolis control with a certain 'selective' temperature T_{sel}. The model studied in [9] is the same as shown in Figure 3. A strong attraction between any pair of amino acids shifted the conformational ensemble in folding simulations towards compact states. The designed sequences were shown to fold into the target (native) conformation, which in all cases turned out to be the non-degenerate global energy minimum.

An attempt to carry out a rigorous test of design for longer sequences (48-mers) in the HP model without introducing strong overall attraction was not successful: the native conformation was always multiple degenerate. The non-compact decoys often had a lower energy than the target conformation. These results are consistent with the earlier prediction [18] and the presented statistical mechanical analysis.

Introducing non-specific additional attraction to bias the conformational ensemble towards compact conformations dramatically slows down folding making it infeasible to fold longer chains [32,43,44]. Thus, the range of lengths that can be studied using the two amino acid type model is very limited. Such limitation may give rise to some small-size artifacts.

Thus, design using the two amino acid type model cannot be successfully extended to longer chains because of the requirement to restrict the conformational ensemble by compact conformations (see Figure 3).

Using 20 types of amino acid. An obvious solution of this problem is to use a greater number of types of amino acids than two. This was done in [18] in which 20 types of amino acids and Myazawa–Jernigan interaction potentials [22] were used. The folding program was carried out for 20 amino acid type model proteins on a cubic lattice (with a fixed composition corresponding to an 'average' amino acid composition in proteins). The designed sequences of 80-mers folded fast and were stable in their target conformation; no conformations with an energy lower than the energy of the target conformation (for the designed sequence) were encountered. These results provided, for the studied model, an important proof that a design approach based on statistical mechanical theory of protein folding is feasible and is basically correct, for the right model.

Using imprinting. A somewhat different, interesting approach to design was proposed by Grosberg and coworkers [20,42]. This approach is based on the idea of pre-biological evolution by 'imprinting', according to which first macromolecules could have evolved as a result of polymerisation of equilibrated monomers, which could have interacted with substrates at a pre-polymerisation stage. The imprinting design procedure also uses the Monte-Carlo annealing protocol, but in the system of disconnected amino acids. After that the chain is threaded through the 'annealed' configuration of monomers on the lattice, thus creating a sequence. The advantage of this method compared to the design procedure proposed earlier in [9,41] is that it can (in principle) be realized experimentally in an abiotic system. A disadvantage is that sequences obtained by imprinting are considerably less stable in their native conformation and sometimes they may not even have the target conformation as their global energy minimum. The reason for this is that sequence design uses the energy function in which nearest neighbors in a sequence do not interact (their interaction adds a constant to the energy of each conformation and is therefore irrelevant). The imprinting method does not take this factor into account. Thus, when a chain is threaded through the annealed system of monomers it will often connect to strongly interacting nearest neighbors, making them bind covalently and therefore lose their strong attraction for stability in the native state. Despite this difficulty, it was demonstrated that the sequences obtained as a result of the imprinting procedure are often able to fold into their native conformation, corresponding to a global energy minimum [20,42].

The statistical mechanics in sequence space

Several authors have proposed optimization techniques, other than Monte-Carlo, to search sequence space [45,46]. In our opinion, the Monte-Carlo search in sequence space is as efficient as other optimization algorithms (because the landscape is smooth and a multitude of solutions exist). The Monte-Carlo approach is advantageous, however, because it converges to the canonical distribution and its results can therefore be rationalized from the statistical mechanical perspective. This interesting analogy between the statistics in sequence space and several statistical mechanical models were noted in [9,18,41,47]. The Hamiltonian for sequence design (Equation 2; in which coordinates are quenched but the amino acid identity variables σ are allowed to vary) is analogous to the Hamiltonian of the Ising model if there are only two types of amino acids and to the Potts model if there are many types of amino acids. It was pointed out in [9,41] that the Monte-Carlo design procedure converges to the canonical distribution in sequence space. Thus, the statistics of sequences become analogous to the statistics of 'spin configurations' in the equivalent statistical mechanical models because they follows the same Boltzmann law. This analogy is explained in more detail in [41], particularly Table 1, in which the one-to-one correspondence between statistical characteristics of sequence design and the Ising model are listed. (Two amino acid type sequences were considered in [41], but the results are trivially generalizable to the multiple amino acid type models.)

Of the analogies above, the most important is probably the relationship between entropy in statistical mechanical models and 'degeneracy' of the protein code. This analogy allows us to calculate $N(\mathrm{E})$ directly from the Monte-Carlo sequence design simulations. The idea of the calculation is based on the thermodynamic equation that relates the entropy at a given temperature T with the average energy at the same temperature using:

$$\mathrm{S(T)} - \mathrm{S}(\infty) = \frac{\mathrm{E(T)}}{\mathrm{T}} - \int_{\mathrm{T}}^{\infty} \frac{\mathrm{E(t)}}{\mathrm{t}^2}\, d\mathrm{t} \qquad (11)$$

$\mathrm{S}(\infty)$ is the entropy of a system at infinite temperature.

In our case of sequence design, the selective temperature, at which Monte-Carlo design procedure in sequence space is carried out, is the temperature in Equation 11. $\mathrm{S}(\infty)$ corresponds to random sequences without a bias towards any particular structure; $\mathrm{S}(\infty) = \mathrm{N} \ln \mathrm{m_{eff}}$. The results of the calculation are shown in Figure 1 for several proteins with the energy function approximation given by Equation 2; the sequence design simulations for each protein in Figure 1 were carried out keeping the amino acid composition fixed and equal to the amino acid composition of the native sequence for each protein [9,41]. (The related results were presented in a recent publication [47]). The solid line in Figure 1 shows a theoretical estimate given by Equation 3. It is quite clear that the theoretical estimate is in excellent agreement with the simulation results. Furthermore, it is clear from Figure 2 that the sequence entropy is approximately the same for all proteins studied (of course, different sequences fold into different protein structures; it is the number of sequences that is invariant for different proteins). Such invariance is understandable because in this approximation the difference in energy functions (Equation 2) between proteins result from the average coordination number of their amino acids and the connectivity (i.e. which of the spatially proximal amino acids are sequence neighbors.

Although these factors are crucial in determining which sequences actually fold into a given conformation, they are not too specific to give rise to pronounced differences in designability. This result of the analysis of the model with 20 types of amino acids can be compared with the 'designability principle' suggested by Finkelstein *et al.* [48] and further addressed by Tang and coworkers [49]. The analysis presented in Figure 2 differs from that of Finkelstein *et al.* in that we did not impose energetic penalties on certain structural features such as turns, whereas these factors were assumed to be important in [48]. On the other hand, the arguments presented in [48] are the phenomenological ones that assume a certain form of density of states for a particular structure; the justification of such assumptions based on a more microscopic model will be very interesting to obtain.

The designability of a protein conformation

Tang and coworkers [49] used a standard 27-mer model [50] with the form of energy function similar to Equation 2. They carried out exhaustive enumeration of all compact conformations and all 'two-letter' sequences. The designability of a structure was defined in [49] as the number of sequences that have this structure as a unique energy minimum among all compact conformations. Interestingly, Tang and coworkers report that certain structures of compact 27-mers are more 'designable' than others in their model. Furthermore, they infer that the designable structures feature protein-like properties such as secondary structure.

It follows from the present analysis that the issue of designability may indeed be important for the models that feature two types of amino acid because some structures can accommodate their 'best' (lowest energy) sequences with slightly lower energies than other structures. In the situation in which there is no significant gap, this small energy difference between different structures is important: a more designable structure can accommodate its sequences with an energy slightly lower than E_c, whereas less designable structures may have an E_{lowest} that is close to or above E_c. These factors can be seen clearly in Figure 3. For the structure shown, the sequences with the lowest possible energy ($E_{lowest} = -84$) exist. The lower the energy of the native state, the lower the probability that a decoy having the same energy will be found (see above and [10,30]). Correspondingly, there may be many sequences that have the structure shown in Figure 3 as their unique ground state (i.e. this structure may be highly designable). It is clear that the designability of the structure shown in Figure 3 results from the special pattern of bonds on the lattice that make it possible to find a sequence that features complete separation between beads of opposite type (sequence neighbors do not interact). There are many structures, however, that do not have such an 'ideal' pattern of bonds so that even their 'best' sequences still have at least one contact between amino acids of opposite type. For these structures, $E_{lowest} = -82$. For the corresponding sequences the gap is smaller and they are therefore less designable than the structure shown in Figure 3. This is consistent with the observation of Tang and coworkers [49] that more designable structures deliver greater energy gaps.

This analysis implies that the pronounced difference in designability exists for the models in which even the maximal possible gaps are small (i.e. $m_{eff} \geq \gamma$). In this case, every favorable contact is important so that differences between structures (patterns of bonds on the lattice) that allow an extra favorable contact to be gained or lost may make a significant impact on the designability. For many types of amino acid, three-dimensional models in which sequences in a target conformation can have an energy that is considerably below E_c (i.e. $m_{eff} > \gamma$) may be highly designable. Thus, it is important to extend the study of [49] to a multiple amino acid type model. Such an extension is, however, a difficult one: it is computationally very costly to enumerate the multi-letter sequences exhaustively as was done for two-letter sequences by Tang and coworkers [49]. The Monte-Carlo simulations in sequence space may be a reasonable alternative to exhaustive enumeration of sequences. The results presented in Figure 2 show no visible differences in designability for the few protein structures that were used for the analysis.

An important caveat of the Monte-Carlo sequence analysis should be mentioned here. The estimate of the number of sequences in Equation 11 is based on the thermodynamic analogy, which is not precise enough to take into account the sub-dominant (in N) contribution to entropy in sequence space. Thus, although the major (exponential in chain length) contribution to the number of sequences that fold into a given structure (corresponding to the linear in N contribution to sequence entropy), is the same for different proteins, there may be sub-dominant (less than exponential in chain length) contributions, which may give rise to some differences in designability. Whether this is so and if it is, whether this is important for our understanding of protein evolution is a matter of future research.

Beyond design with constant amino acid composition

The approach to the design that uses a Monte-Carlo simulation in sequence space with a fixed amino acid composition [9,20,41] is simple, computationally very efficient and non-heuristic (i.e. it is not limited to any particular model of a protein); hence its appeal.

This approach has certain disadvantages, however, the most important of which are: keeping the amino acid composition fixed eliminates the possibility of finding an optimal (for folding and stability) amino acid composition; the assumption of sequence independence of the partition function in Equation 9 (more precisely the contribution to it from non-native decoys) follows from the mean-field heteropolymer theory ([14,19]; this assumption is valid only at high temperature and the deviations from the mean-field predictions need to be examined); and the lack of reference to the temperature at which the sequence is expected to fold — in the full design problem sequence space optimization of P(T) in Equation 9, both the numerator and denominator depend on temperature and it is possible that at different temperatures it becomes important to optimize different factors. The limitations listed above were partially overcome in a number of publications [36,51–53].

The first limitation (constant amino acid composition) was overcome in [36,54]; the quantity $Z = (E_N - E_{av})/D$ (the so-called Z score; [55]) was optimized in sequence space.

Optimization of the Z score instead of native energy corrected one of problems of the simple approach [9,41] — convergence to homopolymeric sequences unless the amino acid composition is constrained. As a result, the design based on optimization of the Z score was also able to find the optimal composition, which provided the best value for the energy gap.

Recent work

A number of recent papers [51–53] addressed the optimization of the Z score, attempting to estimate better the partition function Z rather than simply assuming it to be sequence independent. In general, this problem is very complicated because an exact solution would require enumeration of conformations after each mutation (to evaluate Z for the new sequence), which makes it computationally very difficult for small chains and totally prohibitive for longer chains of realistic length.

Dual Monte-Carlo simulations

Seno *et al.* [53] attempted to optimize directly P(T) in Equation 9 using dual Monte-Carlo simulations — in sequence and conformational space (the chain growth algorithm was applied for the conformational space simulation). This approach requires considerable computational effort in order to reach the Boltzmann distribution to provide a correct estimate of the partition function Z. Even for shorter chains, such equilibration would require more than 10^5 Monte-Carlo steps and this number grows fast with chain length [56] making the interesting approach proposed by Seno *et al.* [53] very demanding computationally. The apparent advantage of this approach is that it contains direct reference to folding temperature and is rigorous. The disadvantage is that it is computationally very demanding if realistic lengths are employed.

High-temperature approximation

Deutsch and Kurosky (DK; [51]) attempted to estimate the partition function in a high-temperature approximation taking into account the first cumulant only by presenting the partition function Z in the simplest form:

$$F_s = -T \ln Z = \sum_{1 \le i < j \le N} [U(\sigma_i,\sigma_j) \langle \Delta(r_i,r_j) \rangle] \qquad (12)$$

where $\langle\,\rangle$ denotes unbiased averaging over all conformations.

It is quite clear that for compact chains the approach of DK is basically equivalent to the earlier approach in [9], which assumed sequence independence of the partition function. Indeed, in globular polymers $\langle \Delta_{ij} \rangle$, which has the physical meaning of the probability of a contact between monomers *i* and *j* in the full ensemble of conformations, does not depend on *i* and *j* except when these monomers are close to each other along the chain [35,57]. It is clear that setting $\langle \Delta_{ij} \rangle = \text{constant}$ in Equation 12 results in sequence independence of the partition function. In apparent contradiction with the above arguments, DK reported a considerable improvement (for the two-letter HP model) over the results of the previous approach [9]. It is possible that the improvement over the simplest design reported in [51] results from the special property of the cubic lattice that excludes the contacts for which $j - j$ is even. In other words, on a cubic lattice $\langle \Delta_{ij} \rangle \approx \text{constant}$ when $i - j$ is odd and is 0 otherwise. The design in [51] took advantage of this property of the cubic lattice providing a proper distribution of H and P monomers over even or odd sites.

The HP model

It is also worth mentioning that both DK [51] and Seno *et al.* [53] used the HP model to test the results of their design procedures. In both cases, the methodologies are not limited technically to the HP model. As was explained before, the HP model is problematic when studying design and folding. For the two-letter model on the square lattice (as well as on the cubic lattice with average attraction between monomers), $m_{eff} \approx \gamma$ (i.e. it is on the verge of failure). This makes the design results for the HP model unstable and heavily dependent on the details of a model such as lattice type, chain length, 'even–odd' contacts, details of the composition, etc. It is quite possible that some improvements of the design methods over the simplest one suggested in [9] actually solve the problems specific to the gapless HP model. These problems may not exist in more realistic multiple-letter models, where any reasonably compact structure is designable even within the simplest algorithm of [9].

To this end it would be desirable to apply the interesting design methods proposed by DK [51] and Seno *et al.* [53] to a 20 amino acid type model and compare folding rates and stability of sequences designed using various procedures.

Cumulant design of sequences with a high probability to be in the native state

Morrissey and Shakhnovich (MS; [52]) proposed a new design procedure, which seeks sequences having high probability to be in their native state at a given temperature T, P(T). This procedure also employs Monte-Carlo in sequence space; the partition function of the chain Z entering the expression for P(T) in Equation 9 is, however, estimated using the cumulant expansion approximation. This eliminates the need to run simulations in conformational space after each mutation to estimate the partition function [53] and therefore dramatically increases the computational efficiency.

This design procedure was carried out for 20-letter model proteins of various sizes (36-mers and 64-mers) on a cubic lattice and turned out to be quite efficient, yielding sequences that are stable at a selected temperature. Two interesting and unexpected results emerged from this study: first, the folding transition temperature for designed sequences turned out to be highly correlated with the input temperature at which designed sequences were stable in their native conformations; second, the temperature at which the folding rate was the fastest, appeared to be very close to the stability temperature T, which was input in the algorithm. This reflects an important feature of proteins in that the optimum of their folding kinetics is achieved at the conditions when their native state is not extremely stable — a finding fully consistent with the well-known marginal stability of natural proteins. The reason for such a relationship between thermodynamics and kinetics is given partly in a simple theory of folding kinetics presented in [32].

The observed correlation between folding rate and folding temperature generates an interesting prediction that proteins from thermophylic organisms should fold very slowly at normal temperature (~300K), a temperature at which folding of mesophilic proteins is fast. This prediction is supported partly by the observation that some thermophylic proteins (e.g. ribonucleotide reductase from *Thermus x*1 [58]) are most active at high temperature (~90°C) and they retain only marginal activity at room temperature. The implicit assumption made here is that enzymatic activity correlates with foldability. The validity of this assumption requires further study.

Interestingly, different features of folding sequences were emphasized in the MS procedure at different input folding temperatures. Sequences that were designed to be stable at high T had a low energy in the native state and a higher dispersion of interaction energies D. In contrast, sequences that were designed to fold at lower temperature had lower D and higher E_N (see Figure 11 of [52]). This result shows that an optimal design strategy may be different for the design of thermostable and mesophile sequences. A possible reason for this was discussed in [52].

Designing longer sequences that fold cooperatively

The theoretical approaches to protein design were based on the results of mean-field heteropolymer theory, which did not take into account inhomogeneity in the distribution of interacting amino acids over the protein structure. This approximation neglects the fact that some parts of the protein (e.g. the interior) may have been stabilized to a greater extent than other parts (e.g. exterior). Lattice simulation showed that this factor may be important for longer proteins giving rise to a 'multidomain' behavior in which the core folds at a higher temperature than the surrounding loops, leading to lower folding cooperativity [59–61]. It was shown [59,61,62] that the existence of domains is correlated with δ, the dispersion of contact energies. Sequences having higher values of δ tend to fold less cooperatively (core first, then loops) whereas sequences with lower values of δ fold as a cooperative unit. An improved design procedure, which optimizes both the Z score and δ was proposed in [62]. This approach makes it possible to design sequences having a desired folding cooperativity.

Evolution-like design of fast-folding sequences

Thermal stability is not the only feature of protein sequences that could be optimized. Another important characteristic is the folding rate. It is of great interest to compare the sequences optimized for stability with the ones optimized for folding rates because it may shed some light on the features of proteins that were optimized in the natural evolution of their sequences. The evolution-like selection of fast-folding sequences was suggested in [63] and further developed in [64]. The idea of the method is conceptually simple and similar to the design that optimizes the stability. Mutations are attempted and only those that make folding faster are accepted (details are given in [63,64]). The algorithm has proven successful yielding many fast-folding sequences. Analysis of the 'database' of emerged sequences showed that they are indeed more thermodynamically stable in their native conformations than random sequences. Interestingly, the Z scores of evolved fast folding sequences were markedly lower than for random sequences, but markedly higher than for sequences that were designed to optimize their Z score (we remind the reader that Z scores are always negative, i.e. 'lower' means 'better' as far as stability is concerned). Despite having a higher Z score, sequences generated by the evolution-like selection procedure folded much faster than sequences designed for higher stability (an order of magnitude at the respective temperatures of fastest folding). This clearly points out the usefulness and limitation of the Z score as a predictor of the folding rate (as well as any other global thermodynamic criterion).

A more detailed analysis of the features of evolved fast-folding sequences showed that their stabilizing interactions were distributed unevenly: acceleration of folding was accompanied by stabilization of a specific fragment of the structure (the 'folding nucleus' [3,4,65–68]), whereas the remaining part of the structure was much less stabilized. In other words, in the evolution-like selection of fast-folding sequences the first few mutations lead to the decrease of Z score accompanied by some acceleration of folding. Further acceleration was achieved after a few subsequent mutations that strengthened a specific set of contacts, the folding nucleus. In the steady state of evolution-like selection in which the folding rate did not change much with mutations, the amino acids at the nucleus positions were remarkably conserved in contrast to other positions in which mutations were frequent.

A similar approach was taken by Ebeling and Nadler [69] in their interesting study of two-dimensional protein models. They pointed out that in their model the energy optimization does not always give the desired results and additional optimization of folding rates may be required to find folding sequences. This conclusion is consistent with the theoretical views presented in this review (see e.g. *Lesson 4*): two-dimensional models behave very differently and the results obtained with these models cannot be compared directly with the results from three-dimensional models. To understand better the differences between two-dimensional models and three-dimensional models it is clearly of interest to study the features of sequences selected for fast folding in [69].

Lessons for folding

The best and most objective criterion of success in protein design the is folding of designed sequences *in vitro*, *in vivo* or *in silica*. Clearly, certain features of the folding phenomenon depend crucially on how the sequences were designed or selected. In particular, sequences that have a large energy gap $E_N - E_c$ fold cooperatively (first-order like). In contrast, weakly designed or random heteropolymers that do not have such a large gap have a non-cooperative folding transition [1,14,19]. Other examples show that features such as on-pathway [59] and off-pathway [54,70] intermediates may be designed 'in' or 'out' by proper sequence selection. For example, the folding dynamics for two sequences designed to fold into the same 36-mer conformation using different design strategies were compared in [54]. The first sequence, Seq1, was designed by optimizing the Z score (at a variable amino acid composition) whereas the second sequence, Seq2, was generated using the original approach [9] that minimizes the native state energy at constant amino acid composition. It was shown that Seq1, which was obtained by optimizing the Z-score, folded faster, more cooperatively and was more stable in the native state than Seq2. The transition for Seq1 followed the two-state scenario both in thermodynamics and kinetics. An equilibrium intermediate and a trapped kinetic intermediate (similar to the equilibrium intermediate) were found for Seq2.

Because both thermodynamics and kinetics are derived from the properties of the energy landscape, there is an established relationship between them (see e.g. [71]). To this end, care should be taken in comparing the results of folding simulations for different models in which sequences were designed differently. Such comparison is possible only if equilibrium behavior of two models are similar. For example, recent studies [72] showed that the folding transition in some off-lattice models is non-cooperative in contrast to lattice models and experiment [18, 73,74]. This fact rules out the nucleation mechanism for the model of [75]. Correspondingly, it may not be very insightful to compare the cooperative kinetics of real proteins and lattice model proteins with the non-cooperative kinetics in the off-lattice model studied in [72,75,76].

The theoretical developments in protein design stimulated interesting experimental studies including design with reduced or simplified alphabets to address the issue of a 'minimalistic' protein sequence (i.e. what is the minimum number of amino acid types that make it possible to design stable folding sequences). Hecht and coworkers [77] designed and synthesized sequences based on the 'two amino acid type' assumption that the distribution of hydrophobic amino acids is the most important determinant of the structure. Although such designed proteins were compact and belonged to the expected (helical) secondary structure class, their folding into a unique structure and their cooperativity has not been fully established. In a recent elegant study by Baker and coworkers [7] the phage display technique was employed to seek minimalistic sequences that fold into the structure of a small protein, SH3, as judged by its activity. Baker and coworkers come to the conclusion that a six amino acid alphabet is generally sufficient for protein design, with an important exception of a few sites for which simplification was not possible. One possibility is that these sites are related to function; another possibility is that they participate in the unique folding nucleus. Future studies will clarify this important issue.

Concluding remarks

One of the main points of this review is that a better understanding of protein folding (at least in the realm of simple models) is of crucial importance to the success of protein design.

Results of statistical-mechanical analysis

The results of statistical mechanical analysis (see Equations 3, 5 and 9 and *Lesson 1*) show that for an appropriate model (for which $m_{eff} > \gamma$) an exponentially (in chain length N) large number of sequences can fold cooperatively into a given structure. This is consistent with the observation that many non-homologous protein sequences can fold into similar conformations [78], the fact that makes the 'bioinformatics' approach to prediction of protein conformation so difficult. From a design perspective, the chance that a designed sequence is identical or even homologous to the native sequence is minimal. Thus, the success of design cannot be measured by relatedness of 'predicted' and native primary structure [46]. When amino acids are categorized into a small number of classes, however, the simplest division being into hydrophobic and polar, the correlation between predicted and real sequences is beyond the noise level [41]. As was noted earlier, the models that have only two types of amino acid essentially fail to fold (unless the ensemble of conformations is very restricted). It is almost tautological to say that design represents a search in sequence space to optimize folding and stability. The straightforward approaches to this problem that directly (from simulations) evaluate the impact of each mutation on folding thermodynamics [53] or kinetics [63,69] are computationally very intensive and at this point are hardly feasible for models other than the simplest lattice models. This calls for a powerful folding criterion that is easy to evaluate without running simulations in conformational space after each mutation. Such a criterion should be a good predictor of folding ability that can be used as a 'scoring function' to be optimized in sequence space. Here, the theory of folding provides a crucial contribution to design, pointing to criteria such as the energy gap, the related Z score and δ, the dispersion of energies of native contacts, and in some cases the stability of the nucleus. Importantly, these criteria correlate with stability and folding rate (in a certain range of temperatures [32,52]) and they have therefore proved very useful for design. A useful folding criterion should be simple and easy to evaluate without intensive searches in conformational space. For example, recently, the so-called σ criterion was proposed to distinguish between fast-folding and slow-folding sequences [11]. Although in essence this criterion is related to the Z score or gap criterion ([4]; A. Dinner, M. Karplus and E.I.S., unpublished observations), its value is not known without the folding simulations. This makes the use of the σ criterion for protein design problematic.

Limitations

Obviously, the folding criteria that are currently used for design have their limitations. In particular, there is evidence that fast folding could have been an important factor in the evolutionary selection of proteins [64,79]. This may call for a criterion that takes the folding kinetics into account more consistently (a step in this direction was outlined in [80]). It is likely that a search for better, simpler folding criteria will remain an important area of research at the interface between protein folding and design.

Another crucial bottleneck in protein design is the lack of knowledge of a potential function that faithfully reproduces protein energetics (i.e. for which the native structure for the native sequence is at the global energy minimum with an energy gap). This direction of research has been extremely active (see e.g. [8,27,81,82]) and is likely to be very active in the future. The major issue here is to find a model that is still feasible to simulate, but which has enough detail to make it possible to derive 'good' folding potentials. It was shown in [27] and by M. Vendruscolo and E. Domany (personal communication) that a simple pairwise-contact potential approximation is too crude to describe real proteins. There is no set of parameters U that provides an energy gap that is sufficient for successful folding simulations of real proteins in the two-body contact approximations of the energetics. It is almost certain that future studies will seek better potentials for more refined models (see e.g. [81,83]) that can be used for reliable design approaches.

A crucial direction of the further study is to bring the progress in theoretical protein design closer to experiment. An important issue that needs to be addressed in applying theoretical models to the design of real proteins is whether the details of sidechain packing are crucial determinants of a protein's structure. Although some original proposals gave affirmative answer to this question [84,85], more recent experimental studies indicated that chain flexibility needs to be taken into account so that many sidechain substitutions can be accommodated by slightly varying the backbone conformations [86,87]. Interesting methods to account for sidechain stereochemistry in sequence selections have been developed [6,88,89] that use the dead-end elimination theorem or Monte-Carlo design, which takes into account the degrees of freedom of the sidechains [90].

An important signature of the maturity of a field is the degree of interaction between theory and experiments. By this criterion, protein design enters its maturity stage and we will undoubtably witness stunning progress in the near future.

Acknowledgements

I am grateful to Alex Gutin, Victor Abkevich, Leo Mirny and Mike Morrissey for their numerous contributions and insights. This work was supported by NIH grant GM52126.

References

1. Karplus, M & Shakhnovich, E. (1992). Theoretical studies of thermodynamics and dynamics. In *Protein Folding*, pp. 127-196. W.H. Freeman and Company, New York.
2. Bryngelson, J., Onuchic, J.N., Socci, N.D. & Wolynes, P. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195.
3. Fersht, A. (1997). Nucleation mechanism of protein folding. *Curr. Opin. Struct. Biol.* **7**, 10-14.
4. Shakhnovich, E.I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
5. Quinn, T., Tweedy, N., Williams, R., Richardson, J. & Richardson, D. (1994). Betadoublet: de novo design, synthesis and characterization of a β-sandwich protein. *Proc. Natl Acad. Sci. USA* **91**, 8747-8751.
6. Dahiyat, B. & Mayo, S. (1997). De novo design: fully automated sequence selection. *Science* **278**, 82-87.
7. Riddle, N.S., *et al.*, & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805-809.
8. Goldstein, R., Luthey-Schulten, Z.A. & Wolynes, V. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA* **89**, 4918-4922.

9. Shakhnovich, E.I. & Gutin, A. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
10. Šali, A., Shakhnovich, E.I. & Karplus, M. (1994). Kinetics of protein folding. A lattice model study for the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636.
11. Klimov, D. & Thirumalai, D. (1996). A criterion which determines foldability of proteins. *Phys. Rev. Lett.* **76**, 4070-4073.
12. Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-230.
13. Ueda, Y., Taketomi, H. & Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pept. Protein Res.* **7**, 445-449.
14. Shakhnovich, E.I. & Gutin, A.M. (1989). Formation of unique structure in polypeptide chains. theoretical investigation with the aid of replica approach. *Biophys. Chem.* **34**, 187-199.
15. Sfatos, C., Gutin, A.M. & Shakhnovich, E.I. (1993). Phase diagram of random copolymers. *Phys. Rev. E* **48**, 465-475.
16. Ramanathan, S. & Shakhnovich, E.I. (1994). Statistical mechanics of proteins with ‘evolutionary selected’ sequences. *Phys. Rev. E* **50**, 1303-1312.
17. Pande, V., Grosberg, A.Y. & Tanaka, T. (1995). Freezing transition of random heteropolymers consisting of arbitrary sets of monomers. *Phys. Rev. E* **51**, 3381-3393.
18. Shakhnovich, E.I. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Lett.* **72**, 3907-3910.
19. Pande, V.S., Grosberg, A.Y. & Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192-3210.
20. Pande, V.S., Grosberg, A.Y. & Tanaka, T. (1994). Folding thermodynamics and kinetics of imprinted renaturable heteropolymers *J. Chem. Phys.* **101**, 8246-8257.
21. Bryngelson, J.D. & Wolynes, P.G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA* **84**, 7524-7528.
22. Myazawa, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.
23. Kolinski, A., Godzik, A. & Skolnick, J. (1993). The general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **98**, 7420-7433.
24. Shakhnovich, E.I., Farztdinov, G.M., Gutin, A.M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice Monte-Carlo simulation. *Phys. Rev. Lett.* **67**, 1665-1667.
25. Socci, N. & Onuchic, J. (1994). Folding kinetics of protein-like heteropolymers. *J. Chem. Phys.* **101**, 1519-1528.
26. Sippl, M. (1990). Calculation of conformational ensemble from potential of mean force. An approach to knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883.
27. Mirny, L. & Shakhnovich, E.I. (1996). How to determine protein folding potential? A new approach to the old problem. *J. Mol. Biol.* **264**, 1164-1169.
28. Finkelstein, A.V., Gutin, A. & Badretdinov, A. (1993). Why are some protein structures so common? *FEBS Lett.* **325**, 23-28.
29. Shakhnovich, E.I. & Gutin, A.M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346**, 773-775.
30. Gutin, A.M. & Shakhnovich, E.I. (1993). Ground state of random copolymers and the discrete random energy model. *J. Chem. Phys.* **98**, 8174-8177.
31. Mezard, M., Parisi, G., & Virasoro, M. (1998). *Spin Glass Theory and Beyond*. World Science, Singapore.
32. Gutin, A., Šali, A., Abkevich, V., Karplus, M. & Shakhnovich, E.I. (1998). Temperature dependence of folding in a simple protein like model: search for glass transition. *J. Chem. Phys.*, in press.
33. Yue, K., Fiedig, K., Thomas, P., Chan, H.S., Shakhnovich, E.I. & Dill, K.A. (1995). A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA* **92**, 325-329.
34. O'Toole, E.M. & Panagiotoupoulos, A.Z. (1993). Effect of sequence and intermolecular interactions on the number and nature of low-energy states of simple model proteins. *J. Chem. Phys.* **98**, 3185-3190.
35. Grosberg, A.Y. & Khohlov, A.R. (1994). *Statistical Mechanics of Macromolecules*. AIP Press, New York.
36. Abkevich, V., Gutin, A. & Shakhnovich, E.I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460-471.
37. Shakhnovich, E.I. & Gutin, A.M. (1989). Frozen states of disordered globular heteropolymers. *J. Phys.* **A22**, 1647-1654.
38. Pande, V., Grosberg, A., Joerg, C. & Tanaka, T. (1996). Is heteropolymer freezing well described by the random energy model? *Phys. Rev. Lett.* **76**, 3987-3990.
39. Govindarajan, S. & Goldstein, R. (1995). Searching for foldable protein structures using optimized energy functions. *Biopolymers* **36**, 43-51.
40. Grosberg, A.Y., Nechaev, S.K. & Shakhnovich, E.I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. (France)* **49**, 2095-2100.
41. Shakhnovich, E.I. & Gutin, A. (1993). A novel approach to design of stable proteins. *Protein Eng.* **6**, 793-800.
42. Pande, V., Grosberg, A.Y. & Tanaka, T. (1994). Thermodynamic procedure to synthesize heteropolymers that can renature to recognize a given target molecule. *Proc. Natl Acad. Sci. USA* **91**, 12976-12979.
43. Gutin, A., Abkevich, V. & Shakhnovich, E.I. (1995). Is burst hydrophobic collapse necessary for rapid folding? *Biochemistry* **34**, 3066-3076.
44. Chung, M., Neuwald, A. & Wilbur, W.J. (1998). A free energy analysis by unfolding applied to 125-mers on a cubic lattice. *Fold. Des.* **3**, 51-65.
45. Jones, D. (1995). Theoretical approaches to designing novel sequences to fit a given fold. *Curr. Opin. Biotechnol.* **6**, 452-459.
46. Koehl, P. & Delarue, M. (1996). Mean-field minimisation methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **6**, 222-226.
47. Saven, J. & Wolynes, P. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. Phys. Chem.* **101**, 8375-8389.
48. Finkelstein, A.V., Gutin, A. & Badretdinov, A. (1995). Why are the same protein folds used to perform different functions? *Proteins* **23**, 142-149.
49. Li, H., Winfreen, N. & Tang, C. (1996). Emergency of preferred structures in a simple model of protein folding. *Science* **273**, 666-669.
50. Shakhnovich, E.I. & Gutin, A.M. (1990). Exhaustive enumeration of all conformations of compact heteropolymers with quenched disordered sequence of links. *J. Chem. Phys.* **93**, 5967-5971.
51. Deutsch, J.M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.* **76**, 323-326.
52. Morrissey, M. & Shakhnovich, E.I. (1996). Design of proteins with selected thermal properties. *Fold. Des.* **1**, 391-406.
53. Seno, F., Vendruscolo, M., Maritan, A. & Banavar, J. (1996). Optimal protein design procedure. *Phys. Rev. Lett.* **77**, 1901-1904.
54. Mirny, L., Abkevich, V. & Shakhnovich, E.I. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of lattice model. *Fold. Des.* **1**, 103-116.
55. Bowie, J.U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-169.
56. Gutin, A., Abkevich, V. & Shakhnovich, E.I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Lett.* **77**, 5433.
57. Shakhnovich, E.I. (1994). *Statistical Mechanics, Protein Structure and Protein-Ligand Interactions*. Plenum Press, New York.
58. Sando, G.N. & Hogenkamp, P.C. (1973). Ribonucleotide reductase from *thermus x*1, a thermophilic organism. *Biochemistry* **12**, 3316-3322.
59. Abkevich, V., Gutin, A. & Shakhnovich, E. (1995). Domains in folding of model proteins. *Protein Sci.* **4**, 1167-1177.
60. Gutin, A., Abkevich, V. & Shakhnovich, E.I. (1998). Cooperativity of protein folding and the random-field Ising model. *Phys. Rev. E*, in press.
61. Panchenko, A., Luthey-Schulten, Z. & Wolynes, P. (1995). Foldons, protein structural modules and exons. *Proc. Natl Acad. Sci. USA* **93**, 2008-2013.
62. Abkevich, V., Gutin, A. & Shakhnovich, E.I. (1996). Improved design of stable and fast-folding proteins. *Fold. Des.* **1**, 221-232.
63. Gutin, A., Abkevich, V. & Shakhnovich, E.I. (1995). Evolution-like selection of fast-folding model proteins. *Proc. Natl Acad. Sci. USA* **92**, 1282-1286.
64. Mirny, L., Abkevich, V. & Shakhnovich, E.I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, in press.
65. Abkevich, V., Gutin, A. & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
66. Itzhaki, L., Otzen, D. & Fersht, A. (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.

67. 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.
68. Shakhnovich, E.I., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98.
69. Ebeling, M. & Nadler, W. (1995). On constructing folding heteropolymers. *Proc. Natl Acad. Sci. USA* **92**, 8798-8802.
70. Abkevich, V., Gutin, A. & Shakhnovich, E.I. (1994). Free energy landscape for protein folding kinetics. Intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6062.
71. Lifshits, E.M. & Pitaevskii, L.P. (1997). *Physical Kinetics*. Pergamon, Oxford.
72. Guo, Z. & Brooks, C. (1997). Thermodynamics of protein folding: a statistical-mechanical study of a small all beta-protein. *Biopolymers* **42**, 745-757.
73. Socci, N. & Onuchic, J. (1995). Kinetics and thermodynamic analysis of protein like heteropolymer: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732-4744.
74. Privalov, P.L. (1996). Intermediate states in protein folding. *J. Mol. Biol.* **258**, 707-725.
75. Guo, Z. & Thirumalai, D. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers* **35**, 137-139.
76. Guo, Z. & Thirumalai, D. (1997). The nucleation collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Fold. Des.* **2**, 377-391.
77. Kamtekar, M., Schiffer, M., Xiong, H., Babik, J. & Hecht, M. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685.
78. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
79. Ladurner, A., Itzhaki, L. & Fersht, A. (1997). Strain in the folding nucleus of chymotrypsin inhibitor 2. *Fold. Des.* **2**, 363-366.
80. Pande, V.S., Grosberg, A.Y., Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a 'new view' needed. *Curr. Opin. Struct. Biol*. **8**, 68-79.
81. Jernigan, R. & Bahar, I. (1996). Structure-derived potentials and folding simulations. *Curr. Opin. Struct. Biol.* **6**, 195-209.
82. Jones, D. & Thornton, J. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210-216.
83. DeWitte, R.S. & Shakhnovich, E.I. (1996). Smog: *de novo* design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **118**, 11733-11744.
84. Ponder, J. & Richards, F.M. (1994). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 5803-5807.
85. Lim, W. & Sauer, R. (1991). The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**, 359-376.
86. Lim, W., Hadel, A., Sauer, R.T. & Richards, F.M. (1994). The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc. Natl Acad. Sci. USA* **91**, 423-427.
87. Baldwin, E., Hajiseyedjavadi, O., Baas, W. & Mathews, B. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of t4 lysozyme. *Science* **262**, 1715-1718.
88. Dahiyat, B. & Mayo, S. (1995). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA* **94**, 10172-10177.
89. De Maeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* **2**, 53-66.
90. Hellinga, H. & Richards, F. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl Acad. Sci. USA* **91**, 5803-5807.