Journal of Computer-Aided Molecular Design, 7 (1993) 397 438
ESCOM



J-CAMD 211

# De novo and inverse folding predictions of protein structure and dynamics

Adam Godzik, Andrzej Kolinski* and Jeffrey Skolnick**

*Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Rd., La Jolla, CA 92037, U.S.A.*

Received 19 October 1992
Accepted 26 February 1993

*Key words:* Inverse folding; Protein folding pathways; Tertiary structure prediction; Lattice protein models; Molten globule intermediates

## SUMMARY

In the last two years, the use of simplified models has facilitated major progress in the globular protein folding problem, viz., the prediction of the three-dimensional (3D) structure of a globular protein from its amino acid sequence. A number of groups have addressed the inverse folding problem where one examines the compatibility of a given sequence with a given (and already determined) structure. A comparison of extant inverse protein-folding algorithms is presented, and methodologies for identifying sequences likely to adopt identical folding topologies, even when they lack sequence homology, are described. Extension to produce structural templates or fingerprints from idealized structures is discussed, and for eight-membered β-barrel proteins, it is shown that idealized fingerprints constructed from simple topology diagrams can correctly identify sequences having the appropriate topology. Furthermore, this inverse folding algorithm is generalized to predict elements of supersecondary structure including β-hairpins, helical hairpins and α/β/α fragments. Then, we describe a very high coordination number lattice model that can predict the 3D structure of a number of globular proteins de novo; i.e. using just the amino acid sequence. Applications to sequences designed by DeGrado and co-workers [Biophys. J., 61 (1992) A265] predict folding intermediates, native states and relative stabilities in accord with experiment. The methodology has also been applied to the four-helix bundle designed by Richardson and co-workers [Science, 249 (1990) 884] and a redesigned monomeric version of a naturally occurring four-helix dimer, rop. Based on comparison to the rop dimer, the simulations predict conformations with rms values of 3-4 Å from native. Furthermore, the de novo algorithms can assess the stability of the folds predicted from the inverse algorithm, while the inverse folding algorithms can assess the quality of the de novo models. Thus, the synergism of the de novo and inverse folding algorithm approaches provides a set of complementary tools that will facilitate further progress on the protein-folding problem.

## INTRODUCTION

The protein-folding problem is deceptively simple to state: given the linear sequence of amino

*On leave of absence from Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.
**To whom correspondence should be addressed.

acids that comprise a given protein, predict the 3D structure of the biologically active, native conformation. With the advent of the human genome project, the necessity of a solution to the protein-folding problem has become especially crucial. The solution to the protein folding problem, apart from its intrinsic interest, would have enormous practical importance. The understanding of the rules of protein folding would enable us to produce proteins with desired properties, such as enzymes with increased thermostability, modified specificity or increased activity [1]. Knowledge of the structure of enzymes and other proteins is a minimal, but necessary prerequisite for rational drug design, and the time-consuming experimental procedures for structure determination delay the development of important drugs [2]. Experimental techniques for determining protein structures have made enormous progress, with more than 1000 protein structures already solved [3,4]. This provides a wealth of information that can be exploited in the development of a successful folding algorithm. While theoretical progress has lagged substantially behind the explosion of experimental information, there has recently been considerable progress on a variety of fronts, as described in detail in this review. A number of algorithms are now available which can match sequences to their corresponding native structure; this is known as the inverse folding problem [5–11]. More recently, de novo folding (where nothing but the amino acid sequence serves as the input) has become possible for real proteins exhibiting simple motifs [12]. What these approaches have in common is that they employ a simplified description of the protein, where many details are suppressed until the final stages of model construction. This review will discuss a number of such simplified, phenomenological approaches to the solution of the protein-folding problem.

Why bother with a simplified description of a protein? And how can one be sure that the simplified representation captures the essence of what a protein is? In other words, why not simply use a full atom model of the protein from the beginning and avoid the apparent complications introduced by a reduced representation? Clearly, full atom models have proven to be extremely useful when investigating processes which depend on local details such as specific enzyme-substrate interactions [13]. Unfortunately, as pointed out by Novotny and co-workers in the mid 1980s [14], extant full atom potentials cannot tell if a protein is correctly folded. Possibly, this reflects available simulation times which are too small or an incomplete treatment of solvation effects, but whatever the physical origin, the ability to identify a correctly from an incorrectly folded protein is a minimal requirement for solving the folding problem. The problem of obtaining good potentials is elaborated on below. Here, we further note that even if one possessed a good full atom potential, there is the multiple-minima problem that must be surmounted. If one uses a full atom description and attempts to sample a reasonable number of conformations, it soon becomes apparent that the cost becomes computationally prohibitive. For example, long full atom molecular dynamics (MD) simulations run for 100 ns or so, whereas proteins fold in the millisecond to second regime. Thus, the full atom realization should be deferred until one is near the native conformation, after which the insights gained over the last 15 years in the handling of full atom models can be exploited.

Over the years, a number of reduced representations of proteins have been employed [15]. These include α-carbon virtual-bond models [16], α-carbon plus β-carbon models [17], α-carbon plus single-ball representations for the side chains [18], α-carbon plus multiple-ball representations for the side chains [19], and α-carbon plus full heavy-atom side-chain representations [20]. With increasing complexity comes additional computational cost, and care must be exercised to

keep the models as simple as possible and yet retain the ability to reproduce features of globular proteins both in terms of geometric criteria, the thermodynamics and the kinetics. Geometric criteria might include the ability to describe the native conformation at a given level of spatial resolution specified in terms of the coordinate and/or distance root-mean-square (rms) deviation; this might be supplemented by the requirement that the model reproduces the side-chain contact patterns embodied in contact maps. Thermodynamic criteria include the requirement that where appropriate the two-state model be satisfied, whereas for kinetics, the rate-determining step in folding should involde a compact intermediate which is closer to the native than the denatured state.

Having decided on a given reduced representation, then a choice must be made to restrict the model to a lattice or allow a continuous description. Clearly, in a continuous description, one need not worry about additional errors introduced by the discretization of space. On the other hand, lattices offer a number of advantages. By formulation, they a priori reduce the number of degrees of freedom that must be searched. In addition, they create a smoother free-energy landscape than exists in the corresponding off-lattice model. There are also direct computational advantages; especially because the local geometric details are fixed, many quantities can be precalculated. The net result of these advantages is that discretized models of proteins offer a speed-up of about a factor of 100 relative to the off-lattice case. However, care must be exercised to ensure that the lattice is sufficiently versatile, that the underlying effects of lattice symmetries are minor and that for the study of dynamic processes, elements of secondary, supersecondary or tertiary structure are not artifically frozen in space due to an incomplete set of local conformational rearrangements [21].

Any model realization, no matter how geometrically exact, still requires a potential energy function which allows one to assess the relative stability of a set of states. As described above, standard full atom potentials cannot differentiate the native from incorrect folds. While it might prove possible to calculate the requisite potentials of mean force from first principles [22], in practice such calculations have proven to be prohibitively expensive. Thus, empirical or semiempirical potentials have been employed. Some are based on experimental free energies of transfer from water to another solvent, assumed to be like the protein interior [23]. These transfer free energies are probably most closely related to the burial free energies of the individual amino acids in a generic protein solvent. Specific interactions between pairs or clusters of amino acids are inaccessible from such measurements, and yet these presumably contribute to the stability of a protein as well. To obtain pair potentials as well as other potentials of mean force, structural databases have been used [9,24–27]. Here, one makes the crucial assumption that the number of occurrences of interest is sufficiently large to obey a Boltzmann distribution. Then, the empirical free energy is estimated by taking minus the natural logarithm of the ratio of the observed to the expected number of occurrences, $n_{exp}$, if the distribution were random. The differences between most interaction scales arise because different reference states for $n_{exp}$ are used. If a statistical approach to calculate the potentials of mean force is adopted, care must be taken to ensure that only high-resolution, well-refined structures are employed, and to avoid biasing to a particular fold, only nonhomologous structures should be used. The result is that adequate statistics for the buried/exposed and pair preferences for individual amino acids can be extracted, but higher order correlations are at the fringe of meaningful statistics [28]. A final approach is to use a nonlinear fitting procedure such as a neural network to choose parameters so that a library of sequences is

matched to a library of structures [11]. In all cases, one must demonstrate that the methodology has predictive capability.

We suggest that the following criteria be employed for the assessment of the validity of a given set of phenomenological potentials:

(1) For a given series of mutations on proteins whose structures are known, can it predict their relative stability? One should be aware that statistical potentials appear to be 'too hot'; that is, the magnitudes of the free energy typically are about a factor of 2 too small [29]. This may reflect the fact that a quasichemical approximation is used in their derivation and not all degrees of freedom responsible for the stability of a protein are accounted for.

(2a) Can it match sequences to the corresponding set of correctly folded structures? This is the simplest version of the inverse folding problem which involves the identification of sequences which are compatible with a given structure.

(2b) Of intermediate difficulty is the ability to identify sequences having 20–30% sequence homology, yet the same 3D structure.

(2c) Of greatest difficulty is the ability to identify sequences having no apparent homology but the same 3D structure. Since this approach is important for the validation of what constitutes a protein, as is its practical utility as an intermediate step before the full protein solution is solved, it forms a major focus of this review.

Criteria 1 and 2 assume a static protein representation that cannot be modified. While these provide useful tools for assessing various aspects of protein stability, more general models should allow for mobility in the folded structure and provide a dynamic assessment of protein stability.

(3) A given sequence when placed in the correct structure should remain stable, or at the very least, the overall topology of the fold should remain the same. When the sequence is placed in an incorrect fold (perhaps from other native structures), it should dissolve.

Criteria 1–3 represent variants of necessary stability tests. More generally, a model should be able to reproduce these features associated with globular proteins:

(4) The denatured state should possess a marginal amount of secondary structure.

(5) The thermodynamics of the conformational transition should satisfy a two-state model [30]. That is, under equilibrium folding conditions, the system should be well described as being either unfolded or in the unique native state.

(6) When appropriate, the model should predict the existence of a unique native state. The packing patterns of the native conformation should resemble those seen in globular proteins. As described below, the side chains in particular exhibit specific contact maps which should be recovered.

(7) The rate-determining step in folding should be closer to the native than to the unfolded state. Based on recent experimental evidence, the compact intermediates or molten globules should have a substantial amount of secondary structure characteristic of the native state but nonspecific side-chain contacts [31]. In contrast, in the native conformation, there should be specific, long-lived tertiary contacts. In addition, the compact intermediates should be slightly swollen relative to the native conformation.

(8) No explicit secondary or tertiary structure information should be encoded into the folding algorithm. Rather, the system should be able to mimic the folding process of real proteins de novo; one should start with a random denatured conformation and finish with a native-like folded protein.

Of course, the criteria that the model must be able to reproduce all the essential features of real globular proteins are rather stringent. As described below, various degrees of success have been achieved in this regard. Nevertheless, a major objective of this review is to demonstrate that substantial progress is being made on the protein-folding problem. It should also be pointed out that simplified models are very powerful tools for investigating various aspects of protein behavior [32–35]. One can vary the relative contributions of secondary structural preferences and tertiary interactions to assess which features are necessary to satisfy the above criteria [36,37]. One can examine the role of various kinds of interactions in determining structural uniqueness. For example, toy models employing just two kinds of residues, labeled hydrophobic and hydrophilic, can be explored [38]. Thus, simplified models by design allow one to focus on specific assumptions about proteins and see what the consequences of these assumptions are. After the basic features implied by idealized versions of such toy models are explored, one can attempt to generalize them to capture the essence of specific real proteins as opposed to those features which are generic to all proteins.

The outline of the remainder of this review is as follows. In the next section, we review recent work on the inverse protein-folding problem as applied to the problem of assigning sequences to global folds. Then, promising work on supersecondary prediction by a 3D alignment method and early simplified models of protein structure and dynamics are described. Subsequently, the status of lattice models of proteins is reviewed; this is followed by a description of a complex lattice model that alleviates all the deficiencies of earlier lattice models. Then, the de novo folding of simple designed proteins is described and a hybrid version that combines 3D alignment results with lattice modeling is presented. Finally, we conclude with a discussion on the future outlook for the protein-folding problem.

## THE INVERSE PROTEIN-FOLDING PROBLEM

### *Background*

While the first protein structures were described by investigators as 'chaotic' [39], very soon regularities were noted, and since then, most analyses of protein structure have been devoted to the search for patterns and/or similarities of various types. There are two key observations that form the core of research approaches in this area. First, for evolutionarily related proteins, a significant portion of their sequences is identical. This observation led to the development of sequence analysis methods, which constitute a well-developed scientific field [40]. Second, both local (e.g. helices [41]) and global (e.g., the myoglobin-phycocyanin folds [42]) structures can be similar without any recognizable similarity in the amino acid sequences. Thus, there is substantial degeneracy in the information that encodes for local as well as global conformation. Precisely what this information is has not yet been fully elucidated.

The first attempt to understand the rules which decide the local regularities came from secondary-structure prediction algorithms. The first methods were formulated in the early seventies [43,44] and since then, despite the development of much more sophisticated methods [45], the accuracy of these predictions oscillates around 60–70%. These methods are also inherently one-dimensional; clearly, 3D information is required to achieve additional success.

What determines a global fold was first analyzed from the point of view of sequence. For instance, when a few sequences from one protein group or family are known, it becomes possible

to build 'sequence signatures'. This was done for a number of protein families such as the globins [42,46,47], the immunoglobulins [48] and the blue copper binding proteins [49]. Both this method, a variant of which was later automated [50], and the whole family of multiple alignment methods [51–56], concentrate on several key positions, with more relaxed constraints on others, and then define a sort of generalized sequence (profile) for the whole group. Homology to this generalized sequence can clearly differentiate between all, however weakly, related sequences and all unrelated ones. The importance of each position can usually be rationalized on the basis of the structure. However, the entire approach concentrates on evolutionary relationships between sequences and there is only a weak correlation between position conservation in the profile and the interactions in the structure [57]. It fails for instance to recognize unrelated sequences exhibiting similar structures (like phycocyanins and globins [42]). The existence of such pairs, and even whole groups of sequentially unrelated, but structurally similar proteins [58]), became a standard example of the limitations of homology-based sequence comparisons.

Other methods tried to match sequences not on the basis of amino acid type, but on the basis of their properties [59]. Whole libraries of amino acid properties were constructed [60,61] and as a result, very sensitive methods for comparing two sequences were developed [59]. A similar approach was used to compare a structure to a sequence by building structural templates [56,62]. Finally, some characteristics of the position in the structure (e.g. being buried/unburied [63], being in a given secondary structure [64], or in a more specific 3D environment [5]) were used to define an extended sequence profile. All of these methods remain essentially one-dimensional in the sense that structural information is being used, but it is compressed and mapped into a 1D profile. No provisions are made for compensating mutations, and the only validation of the alignment comes from analysis of its statistical significance. It is also not possible to rebuild the actual 3D structure from the profile. Still, most successful methods from the last group are able to recognize very distant homologies and most known cases of unrelated proteins with similar structures [5].

Another way of asking the same question is to test what sequence changes can be accommodated without destroying the global fold, given a (known) 3D structure of a protein. The pioneering study focused on the packing of the protein interior [65]. Other methods, which in principle could answer the same question, focused on the problem of reconstructing all atoms in the protein structure from the $C^{\alpha}$ coordinates [66–68]. These approaches share the principal disadvantage of all atom-energy-based methods — they are very time-consuming, requiring several hours of supercomputer time for a single structure-sequence pair. In addition, since in all cases the $C^{\alpha}$ backbone was rigidly fixed, they do not permit even a small relaxation of the structure.

More flexibility was allowed in the model of Finkelstein and Reva [69,70], where protein sequences were matched against an idealized structure, with shifts (but not gaps) allowed. For different β-barrel topologies, this model was able to correctly match sequences into their structural classes. However, the interaction pattern in the idealized structures is far from real, with substantial overcounting of interactions in the core and complete neglect of interactions in the turns and on the surface. In this situation, the results are dominated by the difference between the β-strands at the edge of the sheet and those in the middle — the model may easily be invalid for domains within larger proteins. Also, this approach is not easily extendable to other types of structure.

Recently, a number of authors explored the compatibility between the structure and its sequence in the context of an inverse folding problem [6,8–11]. Typically, there are two interrelat-

ed applications of such algorithms. As schematically illustrated in Fig. 1A, the first application is the problem of finding *all sequences* compatible with *a given structure*. Figure 1B displays the related problem for a *particular sequence* of finding in a library of known structures the *particular structure* that fits the sequence the best. To ascertain that the sequence of interest is likely to be compatible with a particular fold in the library of structures, both comparisons should be performed.

The various inverse folding algorithms differ in the way that the energy of a given sequence in a given structure is calculated; in fact, in some cases the 'energy' should be called a fitness function. A topology fingerprint or a template might be based on interactions between α-carbons or β-carbon positions or a set of contacts based on the distance between all the side-chain atoms. For a given position in the chain, some encode the buried/unburied status, the kind of secondary structure at the position itself and/or between the interacting partners. The potentials can be

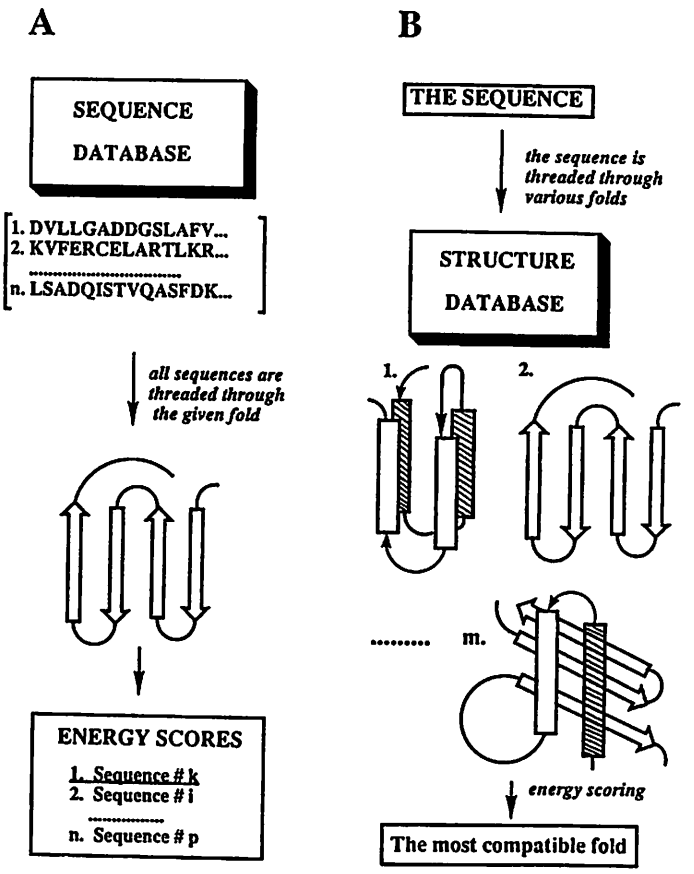


Fig. 1. Schematic illustrations of an inverse folding algorithm designed (A) to find *all sequences* compatible with *a given structure* and (B) when given a *particular sequence*, to find in a library of known structures the *particular structure* that fits the sequence best.

contact based, i.e. defined on the basis of a particular distance cut-off, or they may be distance dependent. The differences between the various approaches are summarized in Table I.

Crucial to the approach is the validity of the potentials. The potentials of mean force that are employed are either based on the statistics of occurrence of the quantity of interest in a structural database or are determined via a fitting procedure based on matching a set of sequences to a database of structures. The validation scheme for both the potentials and the assumed representation of the protein typically consists of attempting to correctly match, for both a training and testing set, a group of N sequences to the library of N structures. We call this the problem of 'mixing and matching'. This may be done with or without allowing for gaps in the structure. Without gaps, a wide variety of potentials can solve the simple problem of matching sequences with their correct structure, but they cannot identify structural twins having insertions [6]. Thus, without the possibility of gaps, the methods are of relatively limited applicability.

The simplest approach is to assume that insertions and deletions only occur in loops or turns and not in elements of secondary structure such as helices or β-sheets [5,10]. If, however, one wishes to allow for insertions and deletions anywhere in the structure, then there are a number of approaches. These differ in how the protein environment is updated to reflect interactions in the sequence of interest relative to those in the original protein which provided the structural fingerprint [6]. Because of its simplicity and apparent generalizability, in what follows we focus on the particular version of the inverse folding algorithm developed by Godzik, Kolinski and Skolnick [7].

*Topology fingerprint approach*

In order to proceed further, one has to decide what constitutes the fingerprint of a given protein fold. We have opted for a description based on the buried/unburied pattern of the residues, and the patterns of contacts between side-chain heavy atoms. For simplicity, two side chains i and j are said to be in contact whenever any pair of atoms in i and j are within 5 Å. This definition will not only permit the construction of a contact map as shown in Fig. 2, but will also allow one to directly focus on the interactions between the residues which involve both two- and three-body terms.

Consistent with the above structural fingerprint description, we proceeded to build a phenomenological interaction scale using a set of 59 high-resolution proteins contained in the Brookhaven

TABLE I
SIMPLIFIED ENERGY CALCULATIONS APPLIED TO THE INVERSE PROTEIN-FOLDING PROBLEM

| Reference | Type of interactions | Distance dependence | Gaps | Ability to build a 3D model | Ability to update environment | Reported examples |
|---|---|---|---|---|---|---|
| 5 | Residue environment | No | Yes | No | No | Several families |
| 6,7 | Contact potential | No | Yes | Yes | Yes | 86 proteins |
| 8 | $C^{\beta}$ | Yes | No | Yes | No | Globins |
| 9 | $C^{\beta}$ | Yes | No | Yes | No | Two families |
| 10 | $C^{\beta}$ | Yes | Loops only | Yes | No | Several families |
| 11 | $C^{\alpha}$ | Two choices | No | Yes | No | 69 proteins |

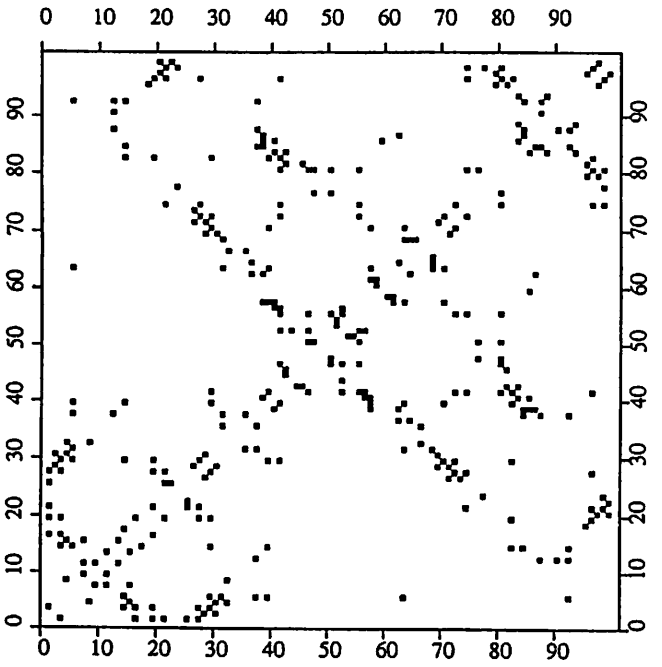


Fig. 2. Contact map of the eight-membered β-barrel protein plastocyanin.

Protein Data Bank. For all classes of terms, the phenomenological free energy is estimated from the negative logarithm of the observed to expected number of occurrences if the distribution were random. Using standard techniques, the accessible surface area for all residues was calculated, and those residues having more than 75% of the total area screened from the solvent were classified as buried. To obtain the expected number of residues, $n_{exp}^1$, if the distribution of buried/unburied residues was random, we calculated the number of residues of a given type multiplied by the average probability of burial for all residues. In a similar manner, the expected number of interacting pairs if the distribution is random, $n_{exp}^2$, is calculated as the total number of interacting pairs multiplied by the fraction of all pairs that interact. In both cases, $n_{exp}^1$ and $n_{exp}^2$ are corrected for protein size, and $n_{exp}^2$ includes a correction for the sizes of the interacting residues. The pair potential is calculated as a conditional probability that the residues are buried or unburied. Finally, we performed a similar analysis on interacting triplets of residues. Surprisingly, the distribution of interacting triplets is far from random. In most of the possible 1330 distinct cases, the statistics are inadequate to draw any conclusions, and for these sets of residues, the three-body interaction energy was set to zero. However, for 74 cases the deviation between the randomly expected and the observed number of residues is sufficiently large that these three-body terms were nonzero. For example, a Cys-Cys-Cys cluster should be observed 377 times in the database as derived from the values of the buried/exposed and two-body interaction energies. In reality, there are only 90 such clusters. Thus, $E_3$(Cys,Cys,Cys) = 1.43 kT. Whether these three-body terms arise from specific interactions of certain clusters of residues or from the simplified protein description which we have adopted is unclear at this time.

Every structure in the database was analyzed and translated into a 'topology fingerprint'. This

'fingerprint' consists of the assignment of each position i as buried/unburied, $\Gamma_i = 1$ or 0, respectively, and a list of all pairs in contact $\{C_{ij}\}$, with $C_{ij} = 1(0)$ if the side chains of residues i and j are (not) in contact. Note that $\{\Gamma_i\}$ and $\{C_{ij}\}$ are defined only in terms of the positions in the structure and are devoid of information about the sequence of the original protein used to define the fingerprint. Thus the total energy is evaluated as follows:

$$E = \sum_i \Gamma_i^{\wedge} E_1(A_i) + \sum_i \sum_{>j} C_{ij}^{\wedge} E_2(A_i,A_j) + \sum_i \sum_{>j} \sum_{>k} C_{ij}^{\wedge} C_{ik}^{\wedge} C_{jk}^{\wedge} E_3(A_i,A_j,A_k) \quad (1)$$

where i, j and k are the positions along the sequence in protein A, and $A_i$, $A_j$ and $A_k$ are the amino acids found at these respective positions. $E_1$, $E_2$ and $E_3$ are the one-, two- and three-body contributions to the total energy*.

One relatively simple test of the interaction scale is to determine whether it can predict the relative stability of a series of mutants. Here, we present the example of different mutants of T4 lysozyme. The free energy of various mutants of lysozyme was calculated using Eq. 1 and then compared to the wild type to give $\Delta\Delta G_{calc}$. Figure 3 presents the values of $\Delta\Delta G_{calc}$ as compared to the corresponding experimental values, $\Delta\Delta G_{exp}$ [71]. Based on the strong correlation of calculation and experiment, we conclude that the interaction scale and the topology fingerprint satisfy Criterion 1.

We then asked the question embodied in Criterion 2a: is the topology fingerprint specific enough that the correct sequence would have the lowest energy when inserted into its own

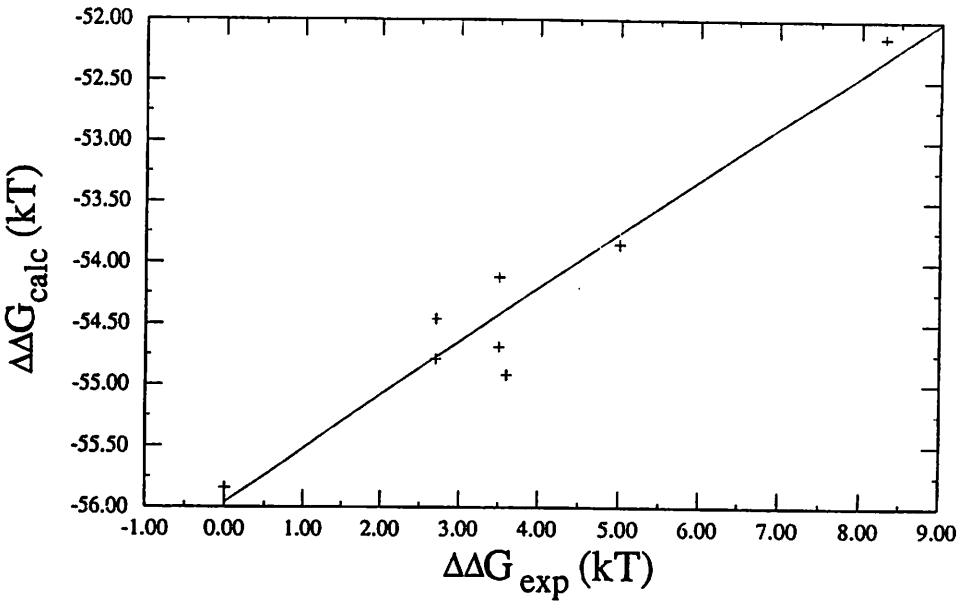


Fig. 3. For T4 lysozyme, a plot of the calculated free energy of a series of mutants relative to the wild-type $\Delta\Delta G_{calc}$ as compared to the corresponding experimental values, $\Delta\Delta G_{exp}$.

*Tables of these parameters are available from the authors via anonymous ftp (file INVERSE on the pub directory of scripps.edu).

structure? Threading the sequence of plastocyanin through all the structural fingerprints yielded the desired answer — not only was the fingerprint of plastocyanin the clear energy minimum, but it was also possible to correctly position the sequence in the fold. The difference between the correct match and the next best answer was more than 30 kT. The same result was obtained for all other sequences in the structural database even when the structures were screened against a sequence database containing almost 22 000 sequences [72]. The procedure was repeated for a set of 30 additional proteins that were not homologous to the set used to derive the energy parameters. All 30 proteins were correctly matched to their respective folds [6]. It has to be noted that this procedure failed when applied to an even wider set of proteins that contained membrane and virus coat proteins, which were mistaken for each other. However, without gaps the method for example cannot identify the structural similarity of azurin, pseudoazurin and plastocyanin (see Fig. 4). Thus, the successful ability to solve the 'mixing and matching' problem, while encouraging, is insufficient to address the full inverse protein-folding problem.

To allow for gaps and insertions, the following approximation was introduced [6]. The energy of every residue in the test sequence is calculated in the environment of the original protein, i.e.

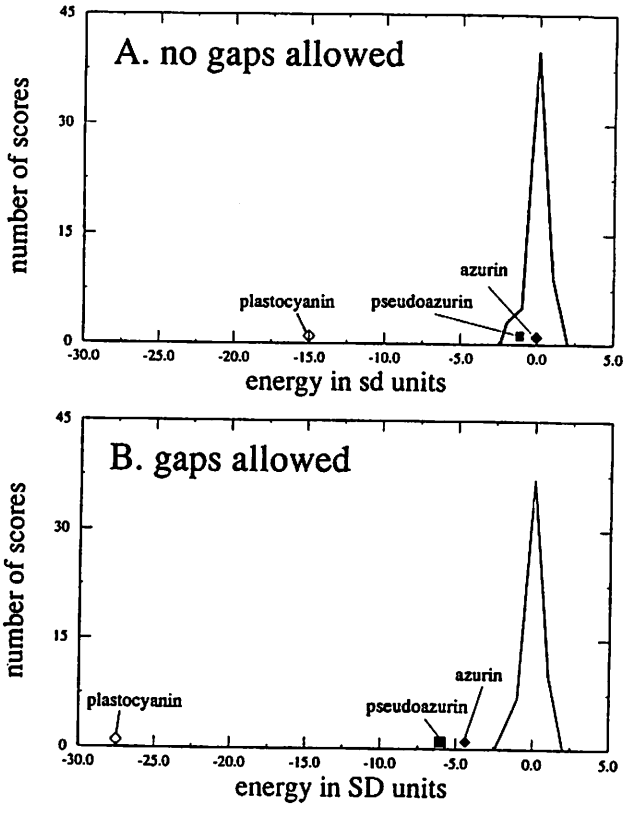


Fig. 4. Distribution of energies of the plastocyanin sequence in different topology fingerprints. (A) Without, and (B) with gaps. The energies of plastocyanin on its own (open diamonds), azurin (solid diamonds), and pseudoazurin (solid squares) are shown in each plot.

the other members of the two- and three-body interaction terms were calculated with residues from the parent sequence corresponding to the fingerprint. That is,

$$E = \sum_i \Gamma^{\wedge}_i E_1(B_i) + \sum_i \sum_{>j} C^{\wedge}_{ij} E_2(B_i,A_j) + \sum_i \sum_{>j} \sum_{>k} C^{\wedge}_{ij} C^{\wedge}_{ik} C^{\wedge}_{jk} E_3(B_i,A_j,A_k) \quad (2)$$

where, as before, i, j and k describe the positions along the sequence, and $\Gamma^{\wedge}_i$ and $C^{\wedge}_{ij}$ are taken from the fingerprint of the protein A, but now $B_i$ is the identity of an amino acid at position i in the sequence which is being investigated. We call this the 'frozen' approximation, since the partners of residue type $B_i$ contributing to the pair interactions ($A_j$) and triplets ($A_j$,$A_k$) are taken from the

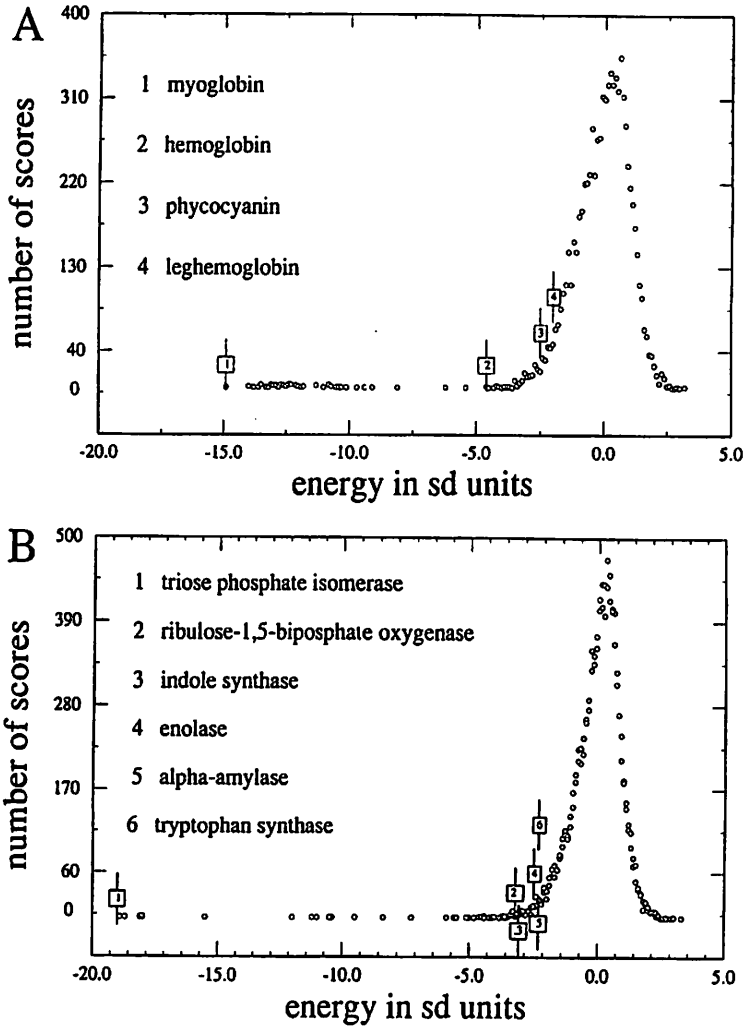


Fig. 5. The distribution of energies of all sequences in the SWISSPROT sequence database [72] for (A) myoglobin and (B) TIM fingerprints versus energy expressed in units of standard deviation from the mean.

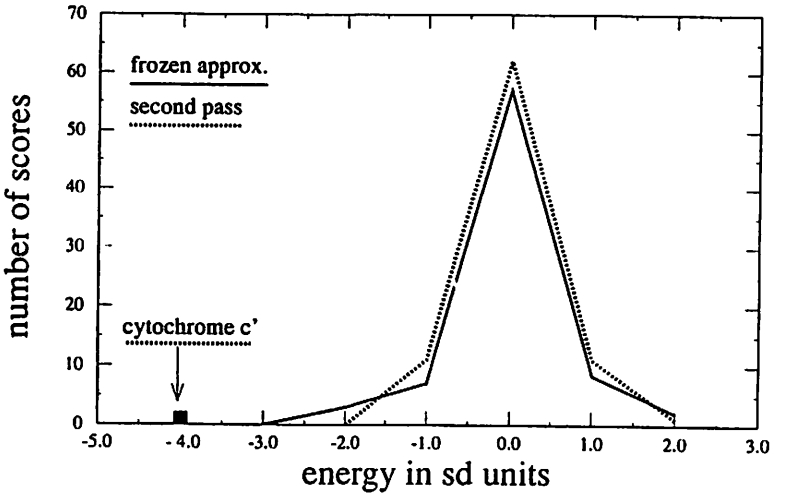


Fig. 6. Distribution of energies of the cytochrome b562 sequence in different topology fingerprints obtained using the frozen-residue environment (solid line) and after updating the environment (dashed line).

original protein, with the rationalization that the interaction environment might not drastically change between equivalent positions in related proteins.

On invoking the 'frozen' approximation, the problem becomes formally equivalent to the traditional alignment of two sequences. The 'frozen' approximation is logically equivalent to the 3D profile method of Bowie et al. [5], that was developed independently and in a different context. Furthermore, not only can we use a structure to define a 'fingerprint', but we can reverse the process to use the 'fingerprint' to build a 3D structure.

An alignment program based on the classical dynamic-programming method [73] was developed. To test the program (and the validity of the approach), the alignment of the azurin sequence and the plastocyanin structural fingerprint was attempted. The best alignment was obtained using values of 2.5 kT for the gap penalty and 0.35 kT for gap extension. Figure 4 shows the results of threading the plastocyanin sequence through the structural library without and with gaps. Plastocyanin [74] and azurin [75] belong to the blue copper binding family and share a similar fold [49], but their sequence similarity is weak. Prior to this work, the correct alignment could only be made by using the most sophisticated and sensitive alignment algorithms. Our algorithm equivalenced 96 positions, with the rms distance between equivalent positions equal to 6.1 Å after superposition. With variations in few places, it is a superset of the best structural alignment [76]; it is not yet possible to identify the conserved core on the basis of the matching procedure alone.

Scanning the sequence database [72] with several other protein fingerprints [6], known examples of proteins with *similar structure, but no sequence homology* were identified. Phycocyanins were selected in the search with the myoglobin fingerprint, and several TIM barrel structures were recognized by the TIM fingerprint. Figures 5A and B show the distribution of scores for myoglobin and TIM fingerprints versus energy expressed in units of standard deviation from the mean. These particular examples were defined as touchstones for a new generation of alignment methods necessary for the solution of the inverse protein-folding problem [77]. Additional alignments

detected using the 'frozen' approximation include the serine proteases, elastase/rat mast cell protease, calcium binding proteins, actinidin and papain, subtilisin and protease K, and the lysozyme-lactalbumin similarity. Such structural alignments have long been well known and lie deep within the 'twilight zone' of 15–20% sequence similarity. In addition, in some cases (see below), similar topological class is identified. These include the structural similarity of the copper binding protein and immunoglobulin folds. Thus, a number of examples of the ability to satisfy Criteria 2b and c have been presented.

A limitation of the frozen approximation will arise if the environments of the test protein and the template protein differ substantially. For example, as clearly shown in the solid line of Fig. 6, if cytochrome b562 (PDB code 256b) is threaded through the library of structures, which includes cytochrome c′ (PDB code 2ccy), no structure scores significantly. This points out the necessity for updating the residue environment from the original fingerprint to that reflecting interactions in the sequence of interest. Following the alignment $N_{A \to B}$, the partners which appear in Eq. 2 are now replaced by the new set of partners $\{B_k\}$ emerging from the first stage of alignment. The energy in the next set of alignments is calculated according to

$$E = \sum_i \Gamma_i^A E_1(B_i) + \sum_i \sum_{>j} C_{ij}^A E_2(B_i, B_{N_{A \to B}(j)}) + \sum_i \sum_{>j} \sum_{>k} C_{ij}^A C_{ik}^A C_{jk}^A E_3(B_i, B_{N_{A \to B}(j)}, B_{N_{A \to B}(k)}) \quad (3)$$

This procedure is iterated until the alignment converges. At this level, one still employs the contact map of the original fingerprint protein. The number of possible partners can either remain the same (if there are no gaps at all in the partner position j and k), or it can be reduced if there are some gaps in the alignment at the positions j and k. Thus, Eq. 3 retains the original contact maps but substitutes different partners at the contact positions.

The results of this approach are demonstrated in the dashed line in Fig. 6 where the 2ccy structure is now identified as having structural similarity with the 256b sequence. Examples of other groups of proteins having similar structure but with non- or weakly homologous sequences include granunocyte macrophage colony stimulating factor with interleukin 4, having a sequence similarity of 21.5%, the N- and C-terminal domains of Rhodenase, and the various dihydrofolate reductases 3dfr, 4dfr and 8dfr.

Having a good putative alignment, it is then important to investigate its stability as a function of the variation in gap creation and propagation terms. As in standard sequence alignment methods, good alignments should be stable over a range of values. By way of illustration, we present in Table 2 the number of aligned residues and the energy of the alignment for the cytochrome b562 sequence in the cytochrome c′ fingerprint. Observe that there is an 'island of stability', i.e. there is a broad range of parameters where the character of the alignment is stable. This provides another check of the reliability of the conjectured structural alignment.

At this juncture, it is clear that there is now a set of almost automated tools that can permit one to address the inverse protein-folding problem on a routine basis. The method is capable of recognizing sequences whose global folds are similar but whose homology is low. As such, it provides a valuable extension to standard sequence-homology techniques. It routinely works in the 'twilight zone' of sequence homology. However, the method requires improvements to provide confidence that it can recognize sequences with similar structure but random homology. The present level of approximation operates in the context of static contact maps but adjusts the possible partners involved in these contacts. To go further, the static contact map approximation

will have to be removed. This will permit the $\{C_{ij}\}$ to be updated to reflect the actual contact map at the current stage of alignment; work in this direction is now in progress and involves the use of lattice and off-lattice side-chain rebuilding algorithms.

## IDEALIZED PROTEIN STRUCTURES

The inverse folding approach by definition is limited to existing protein topologies. The structural prediction will only be valid if an example of the correct topology already exists in the database; otherwise, an incorrect structure can be predicted, simply because it is the 'least worse' among the available choices. This problem might be partially solved, if it were possible to include fingerprints of new folds that are not among known protein structures in the fingerprint library. Such a general description of a new fold can be provided by topological considerations [78], suggested on the basis of extended sequence analysis [79] or simply guessed. However, the description will likely be on the level of simplified topology diagrams. It is therefore desirable to be able to build a realistic topology fingerprint on the basis of such a simplified diagram.

The eight-membered β-barrel structure was chosen for the first application. All topologically possible arrangements of eight β-strands have been previously classified [80]. The set provides an example of the generation of idealized topologies without a priori knowledge of the actual protein structures. We have chosen to examine four different protein structures, plastocyanin, 1pcy; the N-terminal domain of gamma crystallin, 1gcr; prealbumin, 2pab; and a domain from the catabolite gene activator protein, 3gap. As seen in Fig. 7, all four are eight-membered β-barrels, but have different connectivities.

The generic contact patterns describing parallel and antiparallel β-β-strand interactions have

TABLE 2
ALIGNMENT LENGTH AS A FUNCTION OF GAP INTRODUCTION (σ-VERTICAL AXIS) AND GAP EXTENSION (s-HORIZONTAL AXIS) PENALTY FOR THE ALIGNMENT OF CYTOCHROMES c′ AND 562b

| σ\s | 0.00 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 81 | 81 | 85 | 85 | 89 | 89 | 90 | 90 | 90 | 90 | 90 |
| 0.25 | 85 | 86 | 89 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 |
| 0.50 | 88 | 91 | 93 | 93 | 94 | 94 | 96 | 96 | 97 | 97 | 97 |
| 0.75 | 81 | 92 | 93 | 95 | 95 | 97 | 97 | 98 | 98 | 100 | 100 |
| 1.00 | 81 | 84 | 85 | 95 | 99 | 99 | 99 | 102 | 102 | 102 | 102 |
| 1.25 | 79 | 88 | 88 | 92 | 92 | 99 | 101 | 102 | 102 | 102 | 102 |
| 1.50 | 79 | 89 | 89 | 97 | 104 | 105 | 105 | 105 | 105 | 105 | 105 |
| 1.75 | 79 | 89 | 91 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 |
| 2.00 | 73 | 91 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 |
| 2.25 | 79 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 |
| 2.50 | 63 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 |
| 2.75 | 63 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 10 |
| 3.00 | 94 | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 10 | 10 | 10 |
| 3.25 | 94 | 105 | 105 | 105 | 105 | 105 | 105 | 10 | 10 | 10 | 10 |
| 3.50 | 94 | 105 | 105 | 105 | 105 | 10 | 10 | 10 | 10 | 10 | 10 |
| 3.75 | 94 | 105 | 105 | 105 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 4.00 | 94 | 105 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |

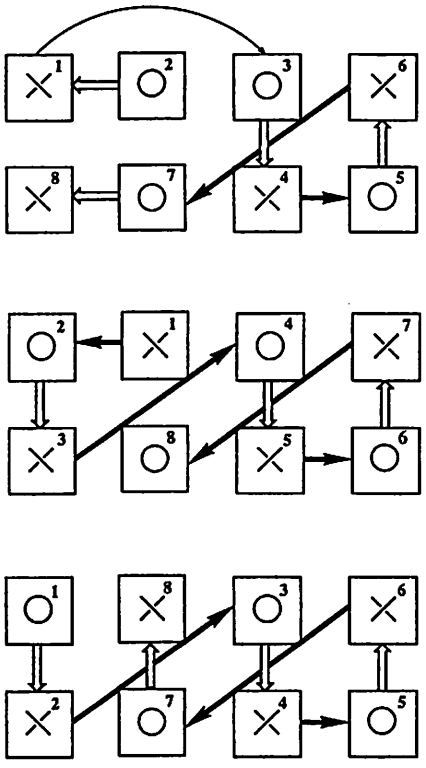

Fig. 7. Connectivity diagrams of the eight-membered β-barrels plastocyanin, the N-terminal domain of gamma crystallin, prealbumin, and a domain from the catabolite gene activator protein, 3gap.

been extracted from the fingerprint database by visual inspection. For each of the four protein sequences and for each of the four topologies, tentative contact maps have been created by including all β-β-interaction regions as read from the idealized diagram. In addition, loops were classified according to their relative position and direction, and interactions between large hydrophobic residues located in close, parallel loops were included in the fingerprint. The fingerprints have been supplemented by assuming the ideal, alternating buried/exposed pattern of β-strands. In the next step, an exhaustive enumeration of all possible loop lengths was performed, and the contact map with the lowest energy was selected. With the exception of the domain from the catabolite gene activator protein, where apparently interactions with the outside of the domain are important to its stability, it was possible to correctly match the sequence to its conjectured contact map.

If the designed and idealized plastocyanin contact map is included in the fingerprint library and the real fingerprint is removed, the idealized fingerprint is identified as the best choice among the structures in the fingerprint library for the plastocyanin sequence. Moreover, the idealized fingerprint recognizes plastocyanin and related sequences among all sequences in the sequence database (see Fig. 8). In other words, starting from the assumption that plastocyanin is an eight-membered β-barrel, it is possible to identify its correct topology.

Obvious generalizations require the removal of the assumption that the number of β-strands is

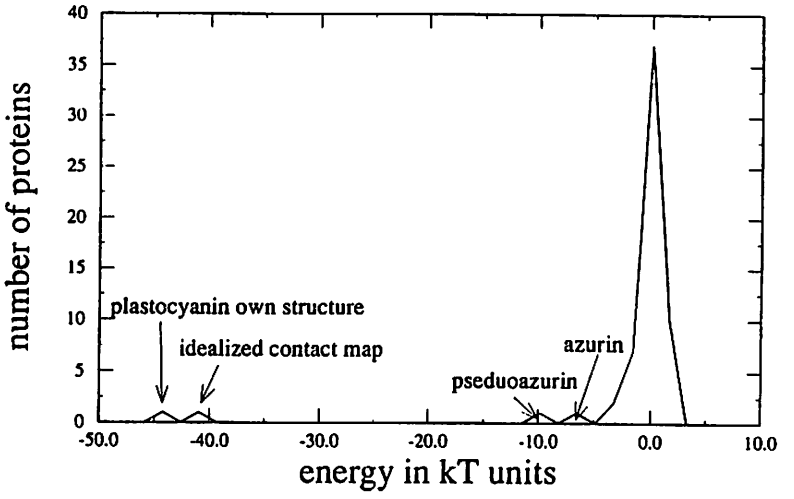


Fig. 8. The distribution of energies of all sequences in the SWISSPROT sequence database [72] for the idealized plastocyanin fingerprint versus energy expressed in units of kT.

known. By examining the density of suboptimal alignments in the structural library, it is possible to determine whether the sequence belongs to an α-, β- or mixed-motif protein. Knowing the dominant type of secondary structure in the protein would permit one to then thread the sequence through all possible folds of the given secondary structural type, and then hopefully to correctly ascertain the most likely topology. Clearly, much more work is required to make this idea a reality.

## SUPERSECONDARY STRUCTURE PREDICTION

If interaction patterns are to be used outside of the inverse folding problem, an important question which must be asked is whether there are smaller fragments of the fingerprint which are specific to their own sequence. If so, it might be possible to predict supersecondary structural elements using an inverse folding algorithm and then assemble the motif using a lattice-based Monte Carlo folding approach. Alternatively, if supersecondary structural elements were identified, this would greatly simplify the topology search discussed above. For example, if β-hairpins could be predicted with high fidelity, one could assign the number of β-strands, $n_b$, in the protein even if one did not know their topological arrangement. Then, an algorithm that exhaustively examines all topologies consistent with $n_b$ strands could be employed. A systematic investigation of this important problem is underway; here, we present some interesting preliminary results.

Three different fragments of full protein contact maps which represent the major supersecondary element motifs were chosen:

(i) An α/β/α fragment from leucine binding protein (residues 121–168 from 21bp).

(ii) A β-hairpin from proteinase B (residues 166–180 from 3sgb).

(iii) An α-hairpin from leghemoglobin (residues 105–146 from 1lh2).

In each case, the structural fingerprint of the fragment (with the same buried/exposed pattern

as in the entire protein) was used to scan the sequence database of nonhomologous proteins with well-refined structures. The energies for all sequence fragments were stored and rank ordered. Qualitatively different results were obtained for each fragment.

The α/β/α fragment possesses the characteristics of a protein; it easily recognized its own sequence and sequences of homologous fragments from other proteins. After this group, separated by large energy gaps, were random scores. This result is consistent with experiments which indicate that α/β/α fragments can be circularly permuted or added to the sequences of existing α/β barrels and retain their identity as autonomous folding units [81].

The α-hairpin had a marked preference for helical hairpins, with its own sequence on the top of the list. There were, however, both α/β and β-hairpins with energies very close to the lowest energy.

The most interesting results have been obtained using the β-hairpin fingerprint. The scores of all protein fragments could be separated into two groups, one with energies lower than or close to zero, and one with energies much higher than zero. β-hairpins of different types constitute about 80% of the first group, and in the top-25 scores only two fragments did not have a β-hairpin structure (both were long fragments of β-strands, with two glycines, or Gly-X-Pro in the middle).

More detailed analysis of the energy profile along a single sequence proved to be even more interesting. As illustrated in Fig. 9 for gamma crystallin, all local minima in the energy profile can be identified within shifts of ±1 residue, with either turns or bends in the DSSP file [82]. In all cases, the template hairpin had no sequence similarity with hairpins identified by the fingerprint. The same qualitative result was obtained for a number of other proteins including plastocyanin and azurin.

There are a number of obvious extensions underway. Fingerprints of the various kinds of hairpins from tight bends to open structures will be extracted. An important question is at what point a supersecondary structural element behaves like a protein. For example, β-hairpins behave

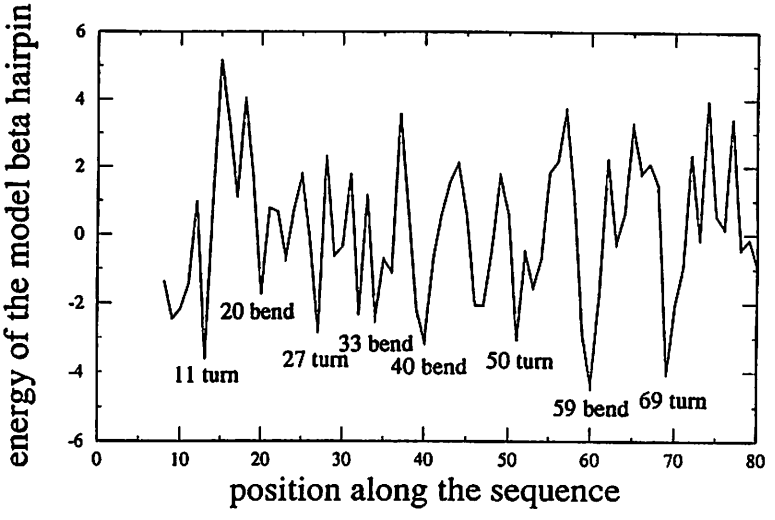


Fig. 9. Energy of fitting sequence fragments of gamma crystallin into the structure of a β-hairpin fragment from proteinase B.

much like secondary elements; many sequences are compatible with the given structure. α/β/α fragments behave as protein — they recognize their own and related sequences. Precisely at what level the specificity of sequence to structure emerges, and what determines it, remain open questions. In particular, the relative contributions of the buried/exposed, pair and three-body terms to sequence specificity require elucidation. Nevertheless, in the context of the present methodology, it is possible to provide answers to these questions.

## SIMPLIFIED MODELS OF PROTEIN STRUCTURE AND DYNAMICS

The first attempts to simulate protein folding started in the mid 1970s and employed a simplified representation of the polypeptide chain consisting of an α-carbon plus a spherical or a β-carbon representation of the side groups [17,83]. This permitted the use of MD to fold small proteins (BPTI). The obtained structures were on the level of 7 Å rms deviation from the X-ray structures [17,84]. The significance of these simulations was questioned after the 'successful' computer folding [85] to the structure of BPTI, using a glycine and alanine heteropolymer, at the same level of accuracy as that obtained when the correct sequence was used.

More recent studies of continuous models employ a similar reduced chain representation. Most of these approaches use a rotational isomeric-states model for the main chain and a single center of interaction located at the center of the side group or at the center of the amino acid. The local chain potential usually accounts for the statistical preferences of the various amino acids for particular structural motifs (i.e. helix, beta or turn). The long-range interactions are usually extracted from the statistics of the PDB X-ray structures [3]. These potentials are sufficient to recover most of the secondary structure [18] and some elements of the global fold of a small protein (crambin). Implementation of an amino-acid-dependent, one-body potential located at the center of mass of the protein allows for an approximate description of some general features of the global fold [86].

The first systematic studies of discrete protein models were performed by Go et al. [87]. The simplest model, consisting of a square lattice chain of 49 residues, addressed the statistical mechanical aspects of the folding and unfolding transition. The dynamics of the transition ('folding pathway') have been studied by means of a dynamic Monte Carlo scheme. The driving force for folding arises from an assumed target contact map and a target set of 'bond angles' down the chain. The effect of nonspecific (nontarget) interactions has also been investigated. Their major findings [88,89] can be summarized as follows:

- Strong short-range interactions consistent with the target fold accelerate folding and make the transition less cooperative.
- Strong long-range interactions, consistent with the target contact map, have less effect on the folding rate; however, the transition becomes more cooperative and the folded structure more stable. Sufficiently strong long-range interactions make the folding transition of the all-or-none type.
- Nonspecific long- and short-range interactions 'diffuse the folding transition' and in the extreme case, lead to locked nonnative states.

The same conclusions [90] hold for the 3D cubic lattice model of lysozyme. They also find that the mobility of the segments corresponding to the surface loops and the ends of the polypeptide chain is larger than for other parts of the model structure.

Krigbaum and Lin [91] performed a Monte Carlo simulation of a bcc (body-centered cubic) lattice model of BPTI. The efficiency of centrosymmetric versus local potentials has been tested. The centrosymmetric potential had a different form for the three classes of amino acids: polar, nonpolar and indifferent. They found that the centrosymmetric potential led to a faster collapse to a dense globular state of the proper volume; however, the resulting structures were rather far from native. Application of the local potential did not improve the resulting structures, while enforcing the correct location of S–S bonds substantially improved the rms.

Dashevskii [92] used a scanning method (similar to the scanning method of Meirovitch [93,94]) for the minimization of the energy of diamond lattice models of PTI and ribonuclease S. Short-range preferences for β-structure and a simple hydrophobicity scale simulated the protein interactions. There is some qualitative similarity between the contact maps of real proteins and those generated by the model.

Since almost all globular proteins are closely packed, with a packing density close to a molecular crystal, one would expect this requirement to grossly reduce the entropy of the system, allowing a fast search for the lowest energy state (states). In this spirit, Covell and Jernigan [95] enumerated compact states of small proteins restricted to various simple lattices. They found that the native state always belonged to the 2% of the states with the lowest energy. The configurational energy of each conformation was calculated based on the number of hydrophobic contacts. Similarly, Shakhnovich and Gutin [96] enumerated all possible chains restricted to a $3 \times 3 \times 3$ lattice cube. The lowest energy compact state of a binary heteropolymer has been found to be in agreement with their theoretical predictions. These computations were extended to cubic lattice polypeptides of the same size [97]. Certain sequences with specific patterns exhibited lower energy. Some other conclusions of this work may, however, be biased due to the lack of ergodicity of the sampling algorithm.

## LATTICE MODELS OF PROTEINS

Lattice models of protein structure and dynamics have proven to be useful at various levels of generalization [32,98]. In several papers, rather simple geometrical representations of protein conformational space were used to analyze the most elemental, and presumably most general, aspects of protein physics. Among various questions which could be addressed are those concerning the cooperativity of protein folding due to the interplay between short- and long-range interactions in the polypeptide chain [88,90,91]. Then, one may examine the requirements for uniqueness of the folded structure and the effect of compactness on the local properties of the model chain [99], which may be associated with the formation of secondary structure in real proteins. Although studies of these models certainly expanded our understanding of the general statistical thermodynamic properties of small, cooperative protein-like systems, their level of abstraction precludes a more explicit description of particular proteins or motifs of protein structure. In this portion of the review, we shall focus on those applications aimed at more accurate representations of protein conformation. Consequently, it is hoped that other properties of protein structure and folding, such as the nature of the molten globule state, the process of chain fixation, and the ability to predict protein structure de novo, can be addressed. For this purpose, we need to evaluate the quality of various discretizations of the protein conformational space.

*Quality of various lattice representations*

The choice of the lattice representation of the protein conformation is usually motivated by the requirement that the long time processes of protein folding be computationally tractable. In most applications, the α-carbon trace of the polypeptide has been modeled by a number of lattice representations. In Table 3, various lattices are briefly compared. They span a wide spectrum of lattices.

The quality of a particular lattice representation can be assessed by various measures [21]. First of all, one may consider the rms deviation of the lattice trace from the X-ray structure of the protein under consideration. However, this measure in and of itself may be somewhat misleading. Even for a quite low overall rms value, the local representation of secondary structure can be very poor indeed. Therefore, one should also consider deviations of the local angular correlations of the lattice approximation from the real chain. Even these requirements are not sufficient for more accurate modeling. In particular, the dependence of the above quantities on the relative orientation of the lattice symmetry axes and the spatial orientation of the protein structure must also be considered. These correlations can be quite strong for some lattices. In other words, a good lattice model should be capable of fitting the geometry of various motifs of secondary structure and various types of proteins at approximately the same level of accuracy, regardless of the particular orientation of the structure with respect to the lattice grid. Table 4 compares the results of the fitting procedure applied to several secondary-structure motifs and to X-ray backbone structures of a few representative globular proteins. Clearly, simple lattices with low coordination number cannot be applied with comparable fidelity to various secondary-structure motifs. Consequently, with increasing protein size, the quality of the fit diminishes. Certainly, the best results are obtained for the high-coordination-number lattice [100]. This is in reality a discretization of the protein conformation, in which the vectors representing the α-carbon–α-carbon virtual bonds belong to a relatively large set of vectors of type [±2, ±1, ±1], [±2, ±1, 0], and [±1, ±1, ±1], including all permutations of the coordinates and the signs of the numbers; we shall refer to this as the hybrid lattice in what follows. Applications employing this fine discretization of protein conformation will be presented in subsequent sections. Here, we will discuss a somewhat simpler model of protein structure and dynamics.

*Monte Carlo folding of simple motifs of protein topology*

In the past, there have been a number of lattice-based studies of the folding of simple motifs representing proteins. Using very simple diamond lattice models and Monte Carlo dynamics, the collapse transition of semiflexible chains of moderate-length chains has the general features of an all-or-none (two-state) transition typical of single-domain globular proteins [101,102]. This extremely simple model appeared to reflect quite well the interplay between short- and long-range (tertiary) interactions seen during protein folding.

Simple diamond lattice chains were successively used for studies of folding of simple protein motifs including four-membered β-barrels [36,103] and four-membered helical bundles of various topologies seen in real proteins [104–106]. The purpose of these studies was to establish the set of minimal requirements for the folding of these model chains to a unique 'native-like' state, starting from an arbitrary random-coil conformation corresponding to the denatured state. In the framework of these simple 3D models, the requirements are the following:

(i) Short-range interactions, or conformational propensities should be on average consistent

TABLE 3
COMPILATION OF PROPERTIES OF VARIOUS LATTICES

| Basis vectors | Lattice unit (Å) | Number of vectors | Name[a] | Reference |
|---|---|---|---|---|
| 100 | 3.80 | 6 | Cubic | 90 |
| 110 | 2.69 | 12 | fcc | 91 |
| 111 | 2.19 | 8 | bcc | 91 |
| 111 | 2.19 | 8(4) | Diamond | 92 |
| 200, 110 | 1.9 | 18 | | 128 |
| 200, 111 | 1.9 | 14 | | 128 |
| 210 | 1.7 | 24 | Chess knight | 110,20 |
| 210, 211 | 1.7 | 48 | | |
| 210, 211, 111 | 1.7 | 56 | Hybrid | 100 |
| 111[b] | 0.81 | 8(27) | Diamond | 129 |

[a] Explanation of names: fcc–face-centered cubic; bcc–body-centered cubic.
[b] In this particular representation the $C^{\alpha}$, C and N backbone atoms are represented on the lattice; therefore, there are $3 \times 3 \times 3$ possibilities to reach from one $C^{\alpha}$ atom to the next in the middle of the chain.

with the expected secondary structure seen in the 'native state'.

(ii) The loops (or turn) regions should have stronger conformational propensities than the rest of the model molecule. Helical proteins required stronger turn propensities than β-type ones.

(iii) In all cases, the pattern of hydrophobic and hydrophilic residues in the sequence should be on average consistent with the expected secondary-structure elements. Then, formation of the hydrophobic core spontaneously emerged.

(iv) Introducing explicit coupling of short-range interactions to tertiary (long-range) interactions makes the folding more cooperative.

It should be pointed out that in all these studies the model system lacked any global, tertiary interaction target potential driving system towards the native-like structure. Nevertheless the folding of these systems was very fast and cooperative [33,34].

A similar model was also used in folding studies of six-membered, Greek-key β-barrels

TABLE 4
THE QUALITY OF VARIOUS LATTICE FITS OF SECONDARY AND TERTIARY STRUCTURE (IN Å)

| Lattice | α-helix | β-strand | αα | ββ | 1 crn | 1 mba | 2 ypi | 1 pcy |
|---|---|---|---|---|---|---|---|---|
| 100 | 1.7 | 2.5 | 1.9 | 2.8 | 2.4 | 2.7 | 3.3 | 4.3 |
| fcc | 1.4 | 0.7 | 1.5 | 1.6 | 1.6 | 1.9 | 2.0 | 3.3 |
| bcc | 1.6 | 1.2 | 2.0 | 1.9 | 1.6 | 2.4 | 2.8 | 2.9 |
| Diamond | 3.4 | 1.2 | 3.8 | 2.3 | 3.2 | 4.0 | 3.8 | 3.8 |
| 200,110 | 1.1 | 1.4 | 1.3 | 1.5 | 1.4 | 1.5 | 2.0 | 1.9 |
| 200,111 | 1.2 | 1.0 | 1.4 | 1.5 | 1.4 | 1.6 | 1.9 | 1.8 |
| 210 | 1.1 | 0.7 | 1.2 | 0.9 | 1.2 | 1.5 | 1.5 | 1.2 |
| Hybrid 211 | 0.7 | 0.6 | 0.9 | 0.8 | 0.8 | 1.0 | 1.0 | 1.0 |
| Backbone | 1.0 | 0.7 | 1.2 | 1.0 | 1.2 | 1.2 | 1.5 | 1.5 |

[107,108]. The folding follows a rather well-defined pathway, and assembly proceeds by a mechanism which can be described as 'on site construction'. The elements of secondary structure do not assemble in isolation and then diffuse together as prefabricated elements. Rather, after forming the first elements of the structure, the remaining portions of the protein assemble by zipping up in a more-or-less correct position with respect to already formed fragments. Of course, at any stage in the folding process, the entire structure or its fragments can dissolve, and the process may recur following a somewhat different string of events. Usually the central core of the barrel forms as an early intermediate, providing scaffolding for the slower assembly of the rest of the model molecule. The transition state has been found to be very close to the 'native state' on the average folding pathway.

While simple lattice models can be used to study highly idealized β-barrel or helical-bundle proteins, they are rather inappropriate for more complex, especially mixed α/β, motifs. For this reason, a chess-knight's walk lattice model has been developed [109]. In the model, the α-carbon trace is represented by a string of [2,1,0]-type vectors on an underlying simple cubic lattice. The excluded volume of the main-chain backbone and the chiral representation of the β-carbon side groups have an accuracy range of 1–2 Å (with respect to the geometry of real polypeptides). This model has been used for detailed studies of the folding trajectories and the stability of the native-like state for several simple motifs, including a model α/β-bundle [37]. The results of these studies show that, at least for rather idealized structures, the folding pathways and basic requirements for stability of the folded state are essentially the same for various lattices. The more detailed, chess-knight lattice models having a simplified representation of the side groups allow for more realistic tertiary interactions. Indeed, in the presence of a very well defined pattern of hydrophobic, hydrophilic and inert (with only the corresponding three values of the hydrophobicity index allowed) residues, these simplified models fold to the proper low-temperature state even when the local conformational propensities are at odds with the final fold.

*Simulation of folding pathways of real proteins*

The chess-knight lattice model of protein conformation has been used in a Monte Carlo dynamics study of the folding of apoplastocyanin [110]. The model polypeptide consisted of an α-carbon backbone representation and side groups of uniform size, about that of a methyl group, for all residues except glycine. Thus, these models are appropriate for the study of the assembly of the overall topology of the native state, but not the late stages associated with chain fixation [111,112]. The full hydrophobicity scale of Miyazawa and Jernigan [26] has been applied to estimate the strength of pairwise, contact-type side-chain interactions. Native as well as nonnative interactions were allowed. The weak short-range interactions, or conformational propensities, were assumed to be in agreement with the α-carbon trace of a 'crumpled' structure of apoplastocyanin. The crumpled structure (6 Å rms from the real 3D structure) has been generated by Monte Carlo relaxation of the original plastocyanin, in order to achieve close packing of the relatively small side groups.

At a proper temperature, the model system always collapsed to the expected unique fold. The folding proceeded down a quite well defined pathway. It usually starts with the assembly of a part of the central β-hairpin, followed by the successive on-site construction of the remainder of the chain. The rate-determining step was the locking of the two last strands of the C-terminus portion of the protein. Folding was relatively fast. Taking into consideration that the number of distinct

conformations of the model chain can be estimated as being of the order of $10^{75}$, the model system certainly finds a way to successively reduce the volume of conformational space that is searched. The Levinthal paradox [113,114] applies to the model system in a similar fashion as it does to real proteins. The requirement of this generation of simulations that the exact local conformation of the protein be known can probably be substantially softened; a relatively accurate prediction of the secondary structure should be sufficient to obtain the unique 3D structure. Such estimates for secondary structural propensities have recently become available from experiment [115].

Somewhat similar simulations [20] have been performed on two large $\alpha/\beta$ proteins; tryptophan synthase and triose phosphate isomerase. However, in this case, the full heavy atom side-chain description has been built into the model, with an average conformation of the side chain assumed. Therefore, the model system had a lower rms deviation from the original structure. However, as above, one can study the pathway of topology assembly, but not the late intermediates of the folding process, including the final fixation of the side-chain positions. Again, a bias towards the proper folds was superimposed on the short-range interactions. The general features of the folding pathways, including the nature of the equilibrium intermediates, have been predicted de novo and are in agreement with experiment [116].

*Validation of the lattice-based approach*

The model studies described in the previous section rely on a dynamic Monte Carlo scheme, which employs an asymmetric Metropolis jump acceptance procedure. The dynamics of the model system are then governed by a stochastic equation of motion, which is solved by the Monte Carlo method. It has been proven that the dynamics of random coil, lattice chains are very similar to the dynamics of off-lattice Rouse chains [117,118] (the correct model of random-coil dynamics in the absence of excluded volume), for which an analytical solution of the equation of motion has been derived. It has also been shown [109] that the introduction of side chains, and consequently local chirality into the model macromolecule, does not introduce any qualitative change in the dynamic properties at high temperature, which corresponds to the denatured state of polypeptides. The updated version of the model [100] exhibits very physical dynamics even at low temperatures with a large amount of secondary structure present. Provided that the elemental moves implemented in the algorithm are local and/or short distance (or there is an appropriate frequency scaling for larger moves) and the scheme is ergodic, then the resulting dynamics should be rather physical. To some extent it should be equivalent to a sparse trajectory obtained from the corresponding Brownian dynamics of closely related models. Consequently, one may expect rather physical trajectories from the Monte Carlo dynamics and physically meaningful folding pathways.

Demonstration of the above-mentioned equivalence was presented in several papers. Using lattice dynamics and Brownian dynamics techniques, Rey and Skolnick [119] studied the folding pathways and equilibrium properties of two similar models of $\alpha$-helical hairpins. They find that the static and dynamic properties of both models are essentially the same, provided that a proper calibration of the interaction parameters and the time scale is performed to account for differences between the fine details of the models. Recent off-lattice Monte Carlo simulations of the assembly process of four helix bundles showed that the pathways [119] and the requirements for unique topology and stability of the native state are essentially the same as in the corresponding lattice simulations. A similar conclusion emerges from comparison of lattice [36,103] and off-

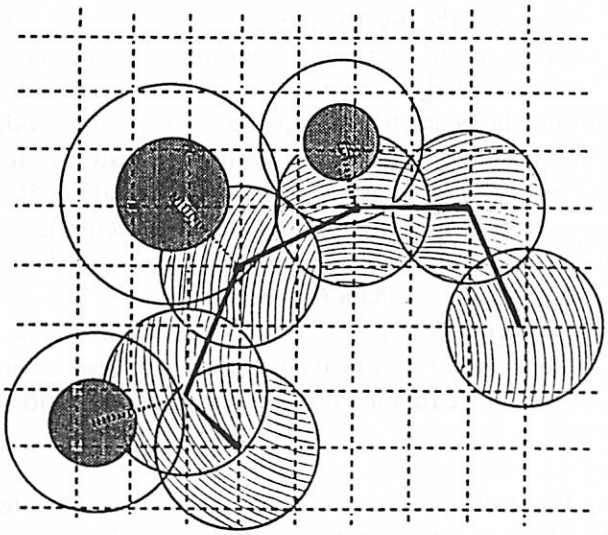

Fig. 10. Schematic drawing of the model chain on the hybrid lattice. The grid shows the underlying simple cubic lattice (spacing 1.7 Å). The shaded spheres represent the hard core of the main chain and the strongly repulsive part of the side groups. The open spheres correspond to the attractive or soft-repulsive part of the side groups. The radii for side groups are approximate, since the distance of interaction is pairwise specific.

lattice [120] studies of otherwise identical models of four-membered β-barrels. Differences between the relevant properties of both models appear to be negligible.

In conclusion, carefully designed lattice models of polypeptides can be safely employed in Monte Carlo studies of protein dynamics, folding intermediates, and native-like structures. It will also be shown below that even phenomena like the molten globule–native-state transition can be simulated using more sophisticated lattice models.

## COMPLEX LATTICE MODEL OF PROTEIN STRUCTURE AND DYNAMICS

In this section, we describe in greater detail a complex discretized model of proteins [12,100], which is capable of reproducing the α-carbon traces of real proteins with an accuracy in the range of 1 Å rms deviation [21]. Here, the use of a large set of side-chain rotamers allows for better packing in the model proteins. Quite complex knowledge-based potentials of mean force make this model very cooperative. At least for smaller proteins, the model allows prediction of 3D structures from just the amino acid sequence. The accuracies of the predictions are estimated to be in the range of 2–5 Å rms (differences between particular folds obtained from separate simulations) depending on the protein size.

### *Geometric representation*

The main-chain backbone is represented by a string of spheres centered on the α-carbons and connected by vectors belonging to the hybrid lattice. The distance of closest approach of two such main-chain segments (except for nearest neighbors along the chain) is assumed to be equal to

4.5 Å. The lengths of the α-carbon–α-carbon virtual bonds are allowed to fluctuate, adopting the values corresponding to 56 vectors from the following set: {[2,1,1],...,[2,1,0],...,[1,1,1]...}. The unit vector of the underlying cubic lattice corresponds to a distance of 1.7 Å. Figure 10 schematically illustrates the geometrical properties of the polypeptide model. The planar valence angle between two consecutive main-chain segments is restricted to the discrete set of values spanning the range seen in real proteins; the lowest value equals 78.5°, and the highest equals 143.1°. The side chains are represented by single spheres. The vector between the α-carbon and the side-chain center of mass depends on the amino acid, the main-chain conformation, and the rotational isomeric state of the side chain (when applicable). The size of the inner (repulsive) and outer spheres of interaction are pair dependent. All characteristics of the side-chain geometry are extracted from a careful statistical analysis of high-resolution 3D structures of globular proteins and are used as a huge rotamer library during Monte Carlo simulations.

*Interaction scheme*

The majority of the potentials of mean force describing interactions between protein segments have been extracted from a statistical analysis of the relevant correlations seen in a structural database of real proteins. Some contributions to the energy of the model system are amino acid (or pair) specific, while other contributions are rather generic, being explicitly independent of the sequence. Because of the complexity of the interaction scheme, a simple descriptive approach is presented below.

*The short-range interactions*

Both a specific and nonspecific part of the short-range interactions are included in the model. The nonspecific (sequence independent) correction to the lattice backbone is introduced to mimic the distribution of distances between four consecutive α-carbons seen in real proteins.

The sequence-specific part consists of a torsional potential acting between side-chain vectors of the form:

$$E_{hm} = \{f(\cos(\Theta_{i,i+2})) + f(\cos(\Theta_{i,i+3})) + f(\cos(\Theta_{i,i+4}))\} \quad (4)$$

where $\Theta_{i,k}$ denotes the angle between the vectors pointing from $i^{th}$ and $k^{th}$ α-carbons to the centers of the corresponding side groups, and f is the statistical potential derived from the structural database, and is a function of identity of both amino acids (the $i^{th}$ and $k^{th}$) involved. The idea is shown schematically in Fig. 11.

*Nonspecific short- and/or long-range interactions*

The hard-core excluded volume of the main chain (vide supra) is slightly exaggerated with respect to the real protein geometry in order to allow for easier side-group packing (consequently, the latter are somewhat reduced in size).

The hydrogen bond-type interaction can also participate in short- as well as long-range interactions (we remind the reader that short-range means interactions between residues close to each other along the chain, while long-range means far apart along the chain, but not necessarily in space). Figure 12A illustrates the lattice definition of the model hydrogen bond. It can be formed between any pair of residues provided that the distance between corresponding α-carbons is

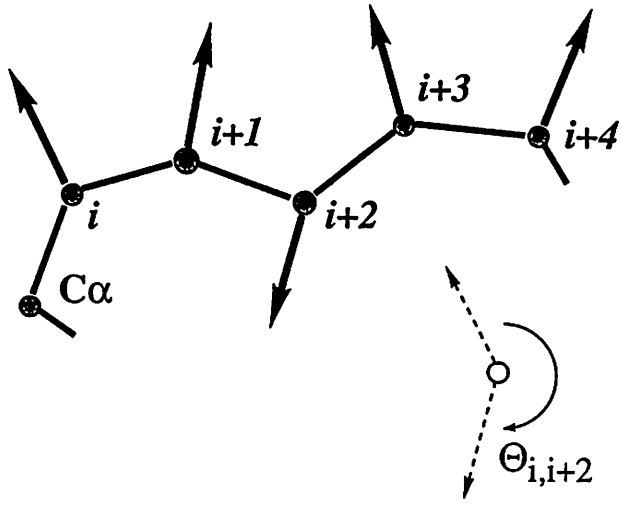


Fig. 11. Illustration of the side-group mutual orientations used for computation of the short-range interactions.

appropriate, and the angular restrictions are satisfied. There is no donor–acceptor asymmetry of the model hydrogen bond; however, the saturation limit of up to two hydrogen bonds to a single backbone residue (except proline, for which the saturation limit is equal to 1) is preserved. Additionally, there is a cooperative contribution to hydrogen bonding. Every time a consecutive pair of hydrogen bonds occurs, the energy of the system decreases. The magnitude of the hydrogen-bond cooperativity parameter is assumed to be the same for helical motifs, β-type structures and loops. Examples are shown in Fig. 12B. The above definition of hydrogen bonds reproduces about 90% of the hydrogen bonds of real protein structures, where the location of the H-bonds is determined by the method of Kabsch and Sander [82].

*The long-range interactions*

There are three different contributions to the long-range interactions. The first is a one-body, amino-acid-specific potential of mean force which depends on the distance of the center of the side

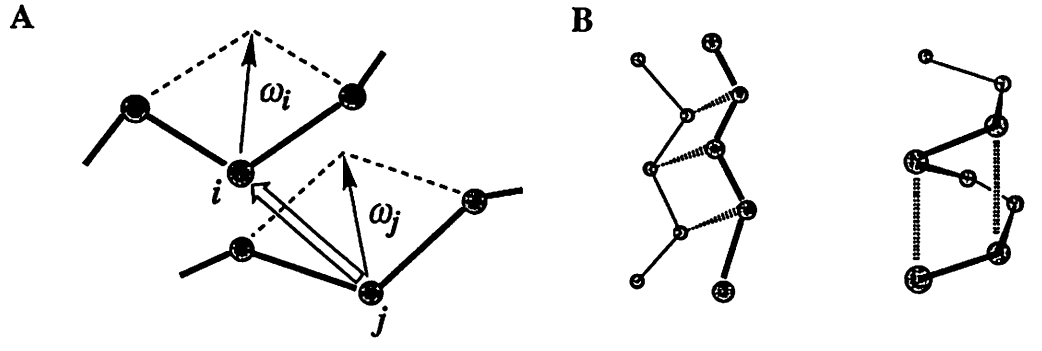


Fig. 12. Schematic illustration of (A) the model hydrogen bonds and (B) two examples of elements of the cooperative network of hydrogen bonds.

group from the center of mass of the model protein chain. It reflects the tendency of some amino acids to be buried in the hydrophobic interior of the globule, while others tend to be buried just under the surface or to be exposed to the surrounding solvent. This term requires the a priori estimation of the radius of gyration of the native state. For single-domain proteins, this is a rather well defined quantity, which can be assumed to depend only on the number of amino acids in the molecule. The small contribution of this part of the potential to the total energy of the folded structure justifies such an approximate treatment.

The second, more important, contribution comes from pairwise interactions between side groups. The two-body potential as well as the pair-dependent contact distances are derived from the statistics of high-resolution 3D structures. As discussed in greater detail below, higher order cluster interaction potentials also seem to be necessary to achieve chain packing patterns typical of native proteins and as well to reproduce the cooperativity of the molten globule-native state transition.

*Monte Carlo dynamics*

A Monte Carlo scheme is used to simulate the long time dynamics of the model proteins. The

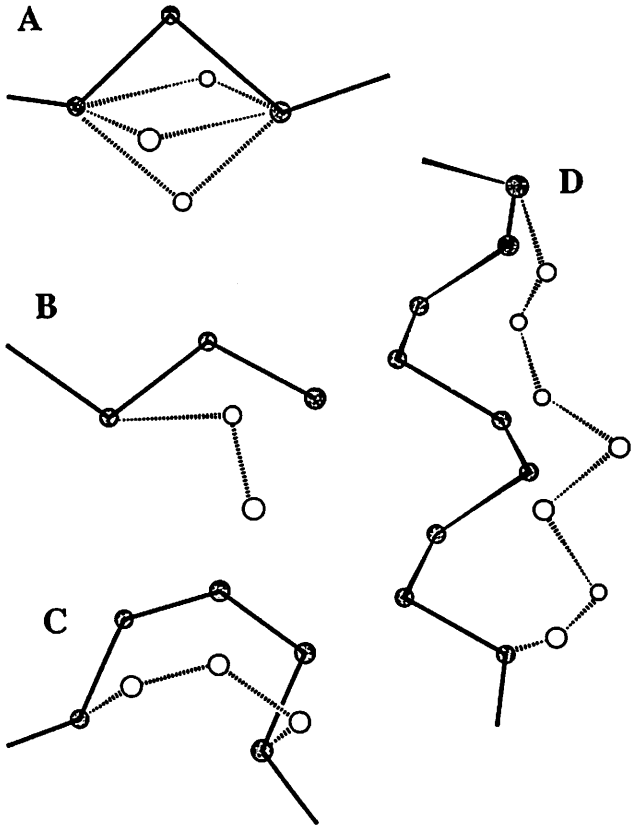


Fig. 13. Some examples of the elemental moves used in the Monte Carlo dynamics. Two-bond spike moves, an example of the end move, and examples of a four-bond move and an eight-bond move are shown. For the sake of clarity, the side groups are omitted.

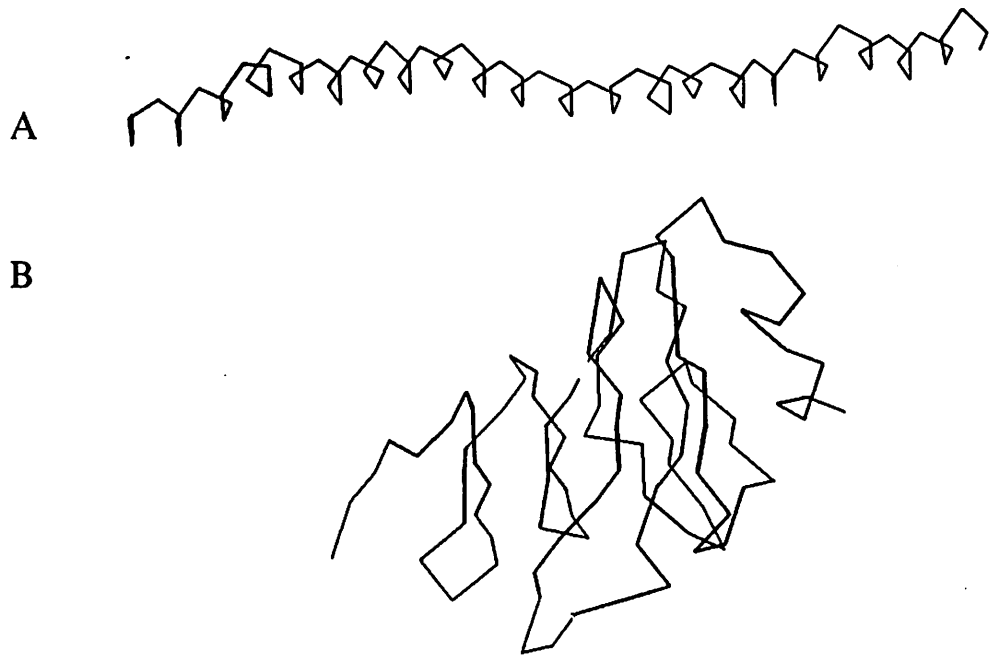


Fig. 14. Typical low-temperature states of (A) the model polyalanine and (B) the model polyvaline. The α-carbon trace is shown in both cases.

rather fine discretization of the conformational space facilitates the use of a large set of local modifications of the chain geometry, some examples of which are displayed in Fig. 13. Each of the above main-chain micromodifications is always accompanied by appropriate arrangements of the side chains. A random change of the side-chain rotamers is also attempted as a separate move in the Monte Carlo scheme.

The simulation algorithm works as follows. First, a type of micromodification, and its particular realization are selected by the random procedure. In the case of violation of geometrical restrictions, the trial conformation is rejected. Otherwise, the change of the energy of the model system is computed, and the new conformation is accepted or rejected according to the Metropolis criterion [121]. A long string of such interations simulates the dynamic thermal equilibrium of the model protein.

*Folding of model homopolypeptides – cooperativity of the transition*

The effect of tertiary versus secondary interactions, and the importance of cooperativity of the hydrogen-bond network on the folding-unfolding transition has been studied for the two extreme cases of homopolypeptides – polyalanine and polyvaline [100]. Alanine exhibits a strong tendency to form helical structures, while valine prefers expanded, β-type structures. In addition, the study of such idealized sequences was expected to provide a straightforward test of the proposed model, before application to the prediction of 3D structures of heteropolypeptides or globular proteins.

In both cases, the same scaling of all the potentials was used. Both polypeptides contained 99 amino acids, a reasonable number for a single-domain globular protein. At high temperature,

both polypeptides adopted random-coil states, with a marginal amount of secondary structure. The dynamics of the random-coil denatured state is very similar to a Rouse-type chain, a model of polymer dynamics in the limit of high solvent viscosity. Both polypeptides undergo a transition from the random-coil state to a globular state; however, the character of the transition and the low-temperature states are different. Polyvaline, being more hydrophobic, undergoes a rather smooth transition to a poorly defined, liquid-like, β-type globule. Polyalanine undergoes a highly cooperative transition to a helical structure which forms at a somewhat lower temperature. Representative snapshots of both systems are shown in Fig. 14. These differences in behavior are caused by different specific short- and long-range interactions. The nonspecific interactions (H-bond potential, excluded volume of the main chain) are the same for both polypeptides. The study of various modifications of the model when particular interactions are artificially suppressed reveals the following:

(i) Hydrogen bonds and the short-range interactions are mostly responsible for the cooperativity seen in the model systems; the helix-coil equilibrium seen for model polyalanine can be well fitted by the Zimm-Bragg theory [122].

(ii) Tertiary interactions alone (when H-bonds and short-range interactions are switched off) lead to collapse to dense amorphous globules. Compactness itself does not induce any appreciable amount of secondary structure.

(iii) The homogeneity of sequences in both polypeptides does not allow for formation of a unique folded state; however, the potential does select for the proper type of secondary structure in both cases.

In conclusion, the model seems to reproduce a number of basic physical properties of globular proteins. Many of the requirements for a proper potential postulated in the introduction have been satisfied by this model. To establish whether or not the remaining requirements can be satisfied requires the de novo folding of real proteins; as described in detail below, the other requirements for a satisfactory protein model can be fulfilled as well.

## DE NOVO FOLDING OF SIMPLE PROTEINS

Before presenting the results of de novo computer folding experiments where, for a given protein, the only specific information provided to the algorithm is the amino acid sequence, we need to discuss an important update to the interaction scheme described above. This involves the inclusion of a four-body potential which introduces cooperativity into the side-chain packing. As a matter of fact, such terms are very much in the spirit of the cooperativity introduced for the H-bond network. While the latter reproduces the cooperativity of secondary-structure formation, the four-body generic correction to the side-chain interactions facilitates a cooperative transition from the liquid-like packing of the side chains in the molten globule state to the well-defined arrangement of the majority of the side chains in the native-like state characteristic of real native proteins.

### *Side-chain packing templates*

Many attempts to predict protein structure from the amino acid sequence alone have failed because the models were incapable of generating the well-defined pattern of side-chain packing exhibited by globular proteins. Even if very detailed pairwise potentials for side-group interac-

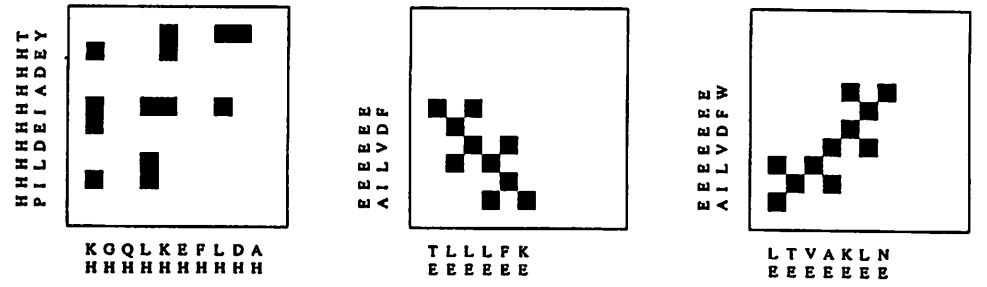


Fig. 15. Ideal patterns of tertiary contacts between two parallel α-helices, two antiparallel and parallel β-strands, respectively. The patterns are part of real protein contact maps.

tions were employed, side-chain fixation or crystallization of the globule interior would probably be extremely difficult, if not impossible to achieve, since the process is somewhat similar to the liquid-solid transition in small-molecule, multicomponent systems. To overcome this difficulty, we again opt for a knowledge-based approach.

First, let us note that on examining the topology fingerprint (the map of contacts between side chains such as in Fig. 2) of real proteins, there are characteristic motifs, or local patterns, which reflect the side-chain packing present in various elements of supersecondary structure. These patterns can be easily identified regardless of the details of the assumed definition of what constitutes a contact between two side-groups. One may use a heavy atom–heavy atom distance criterion or the distance between centers of mass of the side chains. In Fig. 15, three examples of such patterns are shown. The first one is an ideal fragment of the contact pattern for two parallel α-helices. The second one is for two antiparallel, and the third one for two β-strands. In real contact maps, one may see all the features of the patterns shown in Fig. 15 or similar, more or less perturbed, elements. Other elements of secondary and supersecondary structure reflect an analogous level of regularity.

In order to incorporate the possibility of the aforementioned protein-like packing, the cooperativity of the side-chain interactions has been introduced in the following way. Suppose that the side chains of residues i and j are in contact. Then, if another contact (say between residues i + k and j + n) occurs, the system gains an additional energy of stabilization equal to the sum of the two pairwise interactions, provided that the two contacts are members of a subset of the assumed contact map template(s). Otherwise, just the net pairwise interactions are counted. Thus, the tertiary energy of side-chain–side-chain interactions can be written as follows:

$$E_{ter} = \sum_i \sum_j \left\{ E_{ij} + \sum_k \sum_n (\varepsilon_{ij} + \varepsilon_{i+k,j+n}) C_{ij} C_{i+k,j+n} \right\}; \; |k| = |n|, n = \pm 3, \text{ and } \pm 4 \qquad (5)$$

with $C_{ij} = 1$ if residues i and j are in contact (otherwise $C_{ij} = 0$). The correlations with n and k = 4, or n and k = 3 were selected, because they are consistent with both α-helical patterns and various expanded β-type patterns (as well as some features of other secondary-structure motifs) of side-chain–side-chain contacts. The summations should be performed in such a way that any particular contact (the first noncooperative term), or pair of contacts (cooperative contribution) is counted only once. $E_{ij}$ denotes the noncooperative pairwise energy between the two residues which generates liquid-like states and is defined as follows:

$$E_{ij} = \begin{cases} \varepsilon_{i,j} C_{i,j} & ; \text{for } \varepsilon_{i,j} > 0 \\ \varepsilon_{i,j} f / \{(r_{i,j} - d_{i,j})^6 + 1\} & ; \text{for } \varepsilon_{i,j} < 0 \end{cases} \quad (6)$$

with $r_{i,j}$ the distance between two side groups, $d_{i,j}$ the contact distance, and f equal to the square of the cosine between the vectors connecting α-carbons i − 2 and i + 2 and j + 2. The different treatment of the repulsive versus attractive pairs seems to accelerate folding and reflects the fact that the repulsive pairwise interactions have to be of shorter range than the attractive ones, especially for the simplified representation of the side groups employed here.

The particular selection of the templates for binary contacts used here may appear to be somewhat arbitrary. Certainly, it may need further refinement. However, let us note that the templates are consistent with helix–helix as well as with β-strand–β-strand patterns, and enhance some other correlations seen in real proteins. The present realization allows the folding of simple helical proteins to the unique native state with well-defined side-chain packing. While the folding of β-proteins or α/β-proteins has not yet been achieved, the supersecondary structure of these proteins can be predicted with high accuracy. That is, the algorithm generates misfolded structures with proper secondary and supersecondary structures, but with errors in topology of the globular fold. To fold these structures will require longer simulation times and/or further refinements of the method presented here.

*Folding of $\alpha_4$ bundles*

The folding of two proteins whose sequences have been designed by DeGrado et al. [123] has been studied using the methodology described above. Substantial experimental evidence exists that these proteins, or a variant, adopt the four-helix bundle fold. The first sequence (SEQI) is designed to have a hydrophobic core formed by leucine side chains; the second has several leucines substituted by other hydrophobic residues (SEQII) [124]. In particular, the sequences are as follows:

SEQI = $(\text{GELEELLKKLKELLKGPRR})_3$ GELEELLKKLKELLKG

SEQII = GEVEELLKKFKELWKGPRRGEIEELFKKFKELIKGPRRGEVEELLKKFKE-
LWKGPRRGEIEELFKKFKELIKG

For each sequence, a series of random, initial conformations was generated. The folding proceeded by thermal annealing. For SEQI (SEQII), all 24 (14) independent simulation experiments ended with four-helix bundles. A typical pathway, displaying early intermediates is shown in Fig. 16. Assembly of both proteins typically occurs by on-site construction. Usually, as a first noticeable intermediate, the central hairpin forms, starting from the turn, whereas less frequently, one of the terminal hairpins forms first. Then, relatively quickly, a third helix zips up. This three-helix intermediate waits a relatively long time for the last helix to form with proper registration. Of course, at any stage these intermediates often dissolve. Then, a major difference emerges between the two sequences. SEQI, while very stable when assembled, never achieves a well-defined fixation of the side chains within the hydrophobic core. For this sequence, the final stage of the folding simulation has all the properties of a molten globule. There is a compact, slightly swollen conformation with substantial secondary structure. This appears to be in agreement with

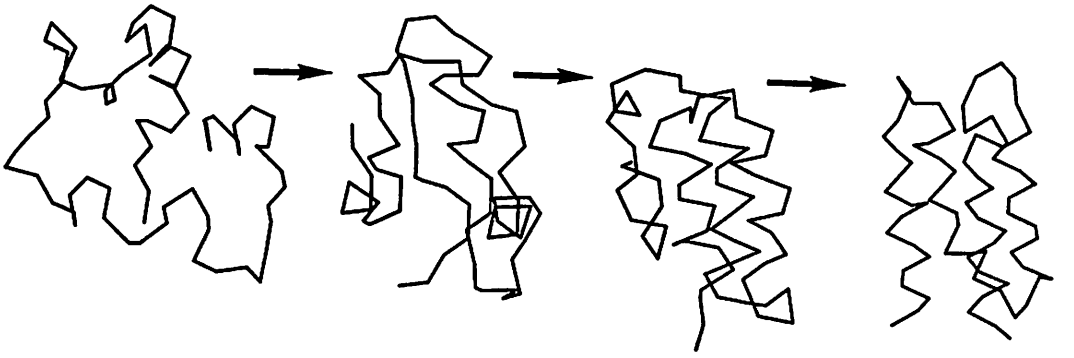

Fig. 16. Representative snapshots of the folding pathway of an $\alpha_4$ bundle with sequence SEQII. The early intermediates are shown. The last snapshot corresponds to the α-carbon trace of the native state of the right-turning bundle.

recent experimental facts [123]. Moreover, the right- and left-turning four-helix bundle topologies were folded with similar probability and are isoenergetic.

After very long simulation times, the second sequence (SEQII) develops a fixed, unique pattern of side-chain packing. The right-turning bundle has a lower energy, and the rms deviation between structures obtained from various starting conformations is smaller, i.e. below 3 Å. Moreover, the chain fixation occurs at low temperature with a relatively large fraction of the side-chain packing pattern (about 75% of the contacts) for this topology reproduced in separate simulations. Therefore, the prediction is that SEQII forms a right-handed, four-α-helix bundle, which is exceptionally stable. The findings for SEQII have to be confirmed experimentally.

Full heavy-atom structures can be built based on the lattice folds, which contain quite accurate information about the main-chain trace, the center of mass positions of the side chains and the pattern of hydrogen bonds. The procedure employs an analytical reconstruction of the full atom backbone [125] including $C_\beta$ atoms. In the next stage, a cut-and-paste procedure can extract the full heavy-atom side chains from a structural database. After Monte Carlo and MD (with the AMBER force field) minimization, the full atom structure can be recovered. An example is given in Fig. 17.

*De novo folding of other helical proteins*

Two other helical proteins have been folded using the method described above. One of them was Felix, the sequence of which has been designed by Richardson et al. [126]. There is experimental evidence that the protein should fold to the left-turning, four-helix bundle. The Monte Carlo simulations in most runs produced this topology; however, the obtained 3D structures seem to be marginally stable. This may reflect real physics since good packing of the side chains of this protein may be very difficult to achieve, and in particular, because the last eight residues of the C-terminal helix seem to be marginally stable, the formation of a fully formed four-helix bundle with attendant stabilization may be partially absent.

Another sequence is the rop monomer which has been obtained by redesigning the naturally occurring rop dimer comprised of two helical hairpins associating into a four-helix bundle [127]. The monomer contains 120 residues and is believed to form a left-turning, four-helix bundle. Preliminary results of Monte Carlo simulations of the folding process indicate that the rop

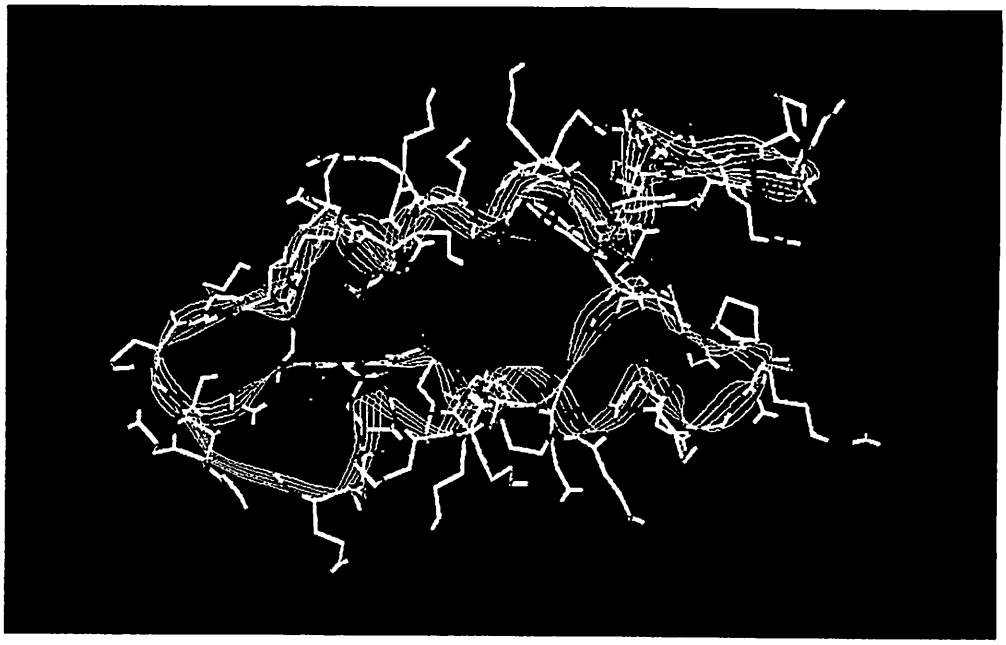

Fig. 17. Right-turning $\alpha_4$ bundle (SEQII) after full heavy-atom reconstruction, and energy minimization (AMBER force field). The hydrophobic residues are shown in purple.

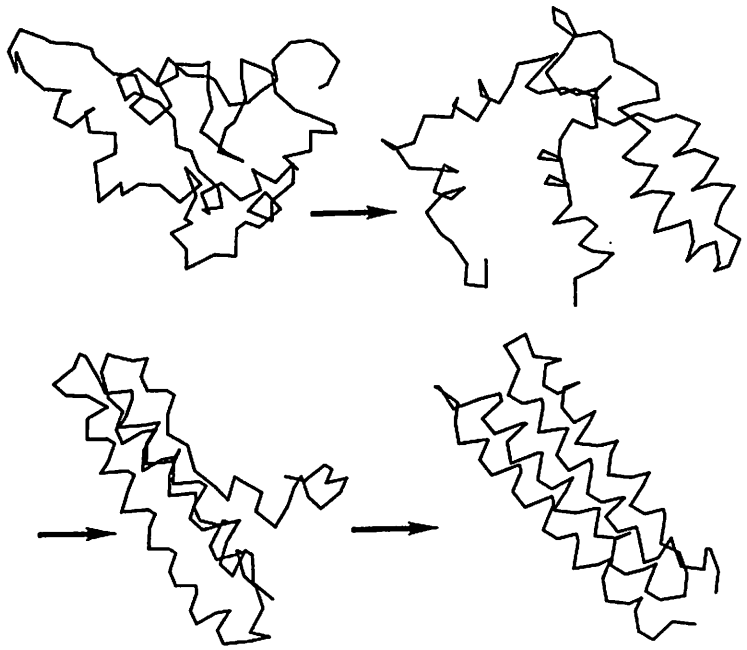

Fig. 18. Representative folding pathway of the rop monomer. The relatively long-living early intermediate is usually a three-helix bundle.

monomer folds to a well-defined, left-turning four-helix bundle. A typical folding pathway is schematically shown in Fig. 18. The usual early intermediate is a three-helix bundle, and the fourth helix assembles on-site after longer times. Then, there is a molten globule intermediate having almost perfect topology (almost always the left-turning bundle is formed during the simulations); however, there is a quite diffuse, liquid-like packing of the side chains. This late intermediate is not very stable and dissolves in most of the simulations. When slightly quenched, the model system adopts a well-defined packing pattern after a long process of side-chain fixation. The energy difference between the molten globule and the native-like state is in the range of 100 kT; this value is substantially larger than that seen for the completely folded protein SEQII. Rop exhibits a much stronger propensity towards helical conformations than the sequences SEQI, SEQII or Felix. Assuming that the rop-monomer 3D structure should be close to that of the rop dimer, and ignoring the modified loop regions, one may estimate the accuracy of the α-carbon traces predicted by simulations to be within the range of 3–4 Å rms deviation from the expected real structure.

## COMBINATION OF INVERSE FOLDING AND LATTICE MODELING ALGORITHMS

*Lattice reconstruction and refinement of 3D alignments*

Once one has identified, using an inverse protein-folding algorithm, which structure is compatible with the sequence of interest, it is then desirable to build a 3D model of the predicted structure. The topology fingerprint approach of Godzik et al. [7] not only provides as output standard sequence alignments between the residues in the fingerprint and the sequence of interest, but it also provides a predicted contact map associated with those pairs of aligned residues. If the sequence homology is high, then one can readily employ extant homology modeling techniques, but if the sequence homology is low, then a number of problems must be faced. First, what should be done about gaps, insertions and deletions? If they occur in loop regions, this is not crucial, but if they occur in elements of secondary structure, modification of the entire fingerprint structure may be required. Ideally, one would like an algorithm that can construct the elements of secondary structure present in insertions and adjust the fold to reflect the interactions which are present in the protein. Another important problem is to assess the stability of the sequence in the conjectured fold. While one cannot rule out the possibility that there is a better arrangement of the chain than those found in the fingerprint library, at the very least the sequence should be stable when inserted into the predicted fold. We describe below preliminary applications of a lattice-based methodology designed to address a number of these points.

By way of illustration, let us examine the problem of identifying the structural alignment of azurin in the plastocyanin fold. Application of the inverse folding methodology equivalenced 96 positions, which translates to an rms of 6.0 Å between these positions if the actual structures are superimposed. We then used the predicted alignment of the azurin sequence into the plastocyanin fingerprint to build the 3D lattice model of azurin. The 30 inserted residues of azurin lack any equivalent positions in plastocyanin, and as indicated above, they initially have random positions subject to chain-connectivity constraints. During the course of the MC simulation, the structure readjusted with more than a 5.8-Å difference from the starting conformation, but the topology remained the same. Following several long MC runs, the rms difference of the model structure from the known azurin crystal structure diminished from 8.3 to 6.2 Å. These rms deviations

should be compared to the inherent resolution of the lattice model without contact templates. For example, the lattice model of plastocyanin is stable, retains the native topology and fluctuates about 6.0 Å from the native structure. This fluctuation can be reduced with the introduction of appropriate contact templates; preliminary indications are that this reduces the fluctuations to 4–5 Å from native.

The simulations described thus far simply demonstrate that the structures are locally stable — perhaps even incorrectly folded structures would not dissolve. To eliminate this possibility, the same protocol was followed for a spurious (in that the global fold is incorrect) but relatively high scoring alignment (in the top 800 sequences) of plastocyanin into the superoxide dismutase monomer. The latter is also an eight-membered, Greek-key β-barrel but has a different sheet packing. The starting conformation was supplied using the topology fingerprint procedure. In contrast to that, when plastocyanin was inserted into the correct fold, the structure slowly diverged and ultimately finished 12.7 Å from the initial conformation. Thus, while the methodology still cannot fold plastocyanin de novo, it can differentiate correctly from incorrectly folded topologies of a given sequence even when the structural class of the topologies is the same.

*Topology fingerprints to verify plausible de novo folds*

The previous section has pointed out one complementary use of lattice models and the inverse folding algorithm, where the former is used to further strengthen the conclusions drawn from the latter. In a reciprocal fashion, one can exploit the inverse folding algorithm to examine the plausibility of de novo predicted conformations. In the de novo folding algorithm described above, one starts with the amino acid sequence and finishes with a putative full atom model of a possible fold. At that point, it is crucial to establish the quality of the predicted conformation. For example, if a topological fingerprint is constructed, how specific is the sequence to the particular fold? If the sequence database is threaded through the predicted structure, does the predicted conformation behave as a real globular protein structure would? That is, does the best-scoring sequence correspond to the sequence which generated the fingerprint fold in the first place, and does it retain similar energetic specificity as those fingerprints constructed from real crystal structures? Furthermore, if the sequence is threaded through the structural fingerprint library which now includes the de novo fold, is it identified as the best fold for the sequence? Thus, the inverse folding algorithm can provide a very useful assessment of the quality of the structures produced by the de novo folding algorithm.

By way of illustration, structural fingerprints were constructed for both right- and left-turning bundles of SEQI and SEQII. For both sequences, based on the threading of the SWISSPROT 15.0 sequence library [72] through each of the right- and left-turning forms, the inverse protein-folding algorithm identifies these as behaving as naturally occurring sequences do. Figure 19 shows the results of threading all sequences through the right-turning SEQII bundle. Thus the first test, demonstrating specificity of sequence to a particular structure, has been passed.

Next, we must demonstrate that the sequence of interest prefers the predicted fold over all other folds in the topology fingerprint library. As shown in Fig. 20, for the right-turning fingerprint of SEQII this is also the case. In fact, the inverse folding algorithm predicts that the right- and left-turning bundles of SEQI are isoenergetic (having energies of −41.2 and −42.9 kT, respectively), in apparent accord with experiment, and that the right-turning SEQII bundle is preferred for the SEQII sequence with an energy of −104 kT as compared to −87 kT for the left-turning,

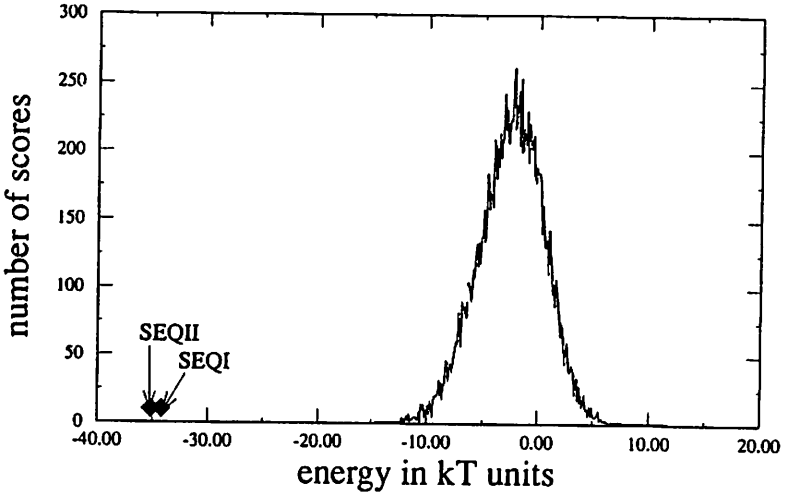


Fig. 19. The distribution of energies of all sequences in the SWISSPROT sequence database [72] for the right-turning fingerprint of SEQII versus energy expressed in units of kT.

four-helix bundle. It is important to note that there is a consonance of the predictions of the on-lattice, off-lattice and inverse folding methodologies with the predictions of the relative stabilities of the right- and left-turning bundles of SEQI and SEQII.

Thus, we conclude that the topology fingerprint approach provides a very powerful adjunct methodology for assessing the relative stability and quality of predicted de novo folds. The fact that the predicted structures behave in an identical manner as protein structures determined by experiment provides confidence that the lattice models have, at least, captured some features of real globular proteins.

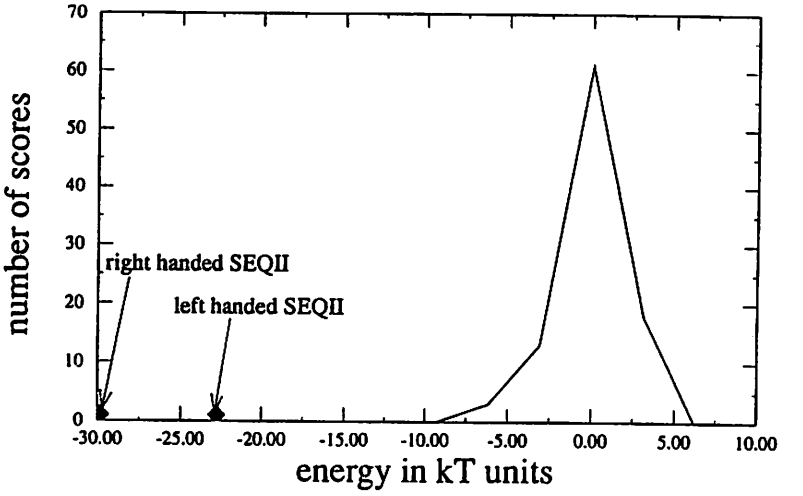


Fig. 20. Distribution of energies of the SEQII sequence in different topology fingerprints including the right- and left-turning four-helix bundle folds predicted for SEQII.

## FUTURE OUTLOOK

In this review, we have focused on the use of simplified models of globular proteins and have demonstrated that such models have widespread applicability. They have proven to be very powerful tools in the treatment of the inverse protein-folding problem, i.e. the determination of sequence–structure compatibility. The inverse folding problem has shown itself to be an extremely useful proving ground for the development of potentials of mean force in particular and for ideas about what determines the specificity of folds in general. While much clearly remains to be done, at the present stage of development it is now possible to identify sequences having apparently random homology and yet which adopt the same global fold. By focusing attention on the interaction patterns present in globular proteins, the inverse folding problem has also been a valuable adjunct in the design of more general de novo models of protein folding. One of the key features of our de novo lattice-based approach, side-chain–side-chain contact templates, was suggested by the patterns of contacts seen in the topology fingerprints. Indeed, generalization of such packing patterns will go hand in hand with the development of fingerprints based on topological considerations.

At this juncture, considerable progress on the protein-folding problem is apparent. For simple topologies, it has proven possible to reproduce both the essential features of protein-folding pathways including the molten globule intermediate states and to fold to unique native structures using as input the amino acid sequence alone. How far this approach can go remains an open question, but a number of directions are immediately apparent. One of the by-products of the current generation of lattice models is a quite reliable predictor of supersecondary structure. Preliminary indications based on the examination of about ten proteins are that the lattice models can predict supersecondary structure with greater than 85% accuracy. This apparent success will be scrutinized more carefully in the near future. Moreover, we will attempt to fold β-proteins such as gamma crystallin and mixed-motif proteins such as flavodoxin. These should establish the range of validity of the present realization of lattice models.

Should these approaches fail, a number of other options are available. First, we shall more fully explore the ability of the inverse folding approach or the lattice-based approach to predict supersecondary structure. Then, a lattice folding algorithm can possibly be employed in a biased search to assemble the various pieces. More generally, since determination of the secondary-structure class of a protein is relatively straightforward, entire fragments of proteins can be used as a structure library which are cut and pasted together on the lattice using empirically determined mixing rules. This might greatly speed up the computer time required for folding. We also envision generalizing the side-chain contact templates to explicitly reflect the backbone conformation of the interaction side chains. For example, if the backbone conformations are helical, then helix–helix templates would be employed. Again, we plan to exploit the synergism of the lattice-based and inverse protein-folding approaches. The various mixing rules for contact templates will be first applied in the inverse folding approach to build idealized fingerprints and then to lattice-based models. Knowledge of how contact maps are constructed will permit us to take idealized topology diagrams and convert them into topology fingerprints. These will greatly increase the size of the topology fingerprint library.

Another promising avenue of investigation is in de novo design of proteins. Here the strategy is as follows. First, one builds a lattice-based model of the native state and verifies that in the

context of the model the side-chain packing is well defined, that is, it is not a molten globule. Next, de novo folding of the designed sequence is attempted. We strive for a sequence that folds relatively quickly. Then, the resulting folded structures are subjected to the inverse algorithm to ensure that the sequence retains the structural specificity of native globular proteins. If the sequence passes these tests, then in the context of these models, it is a worthwile candidate to study experimentally. Experimental studies will prove invaluable in establishing the range of validity of the present approach.

In summary, it is our view that semisimplified models have contributed to substantial progress in the globular protein-folding problem. While it is certainly true that the full protein-folding problem is not yet solved, there are a number of partial solutions. For many, but clearly not all sequences, it is possible to match them to the correct global fold. Based on the rapidity of progress to date, additional advances should be in the offing. Perhaps even more encouraging is, that for both designed and almost naturally occurring sequences, it is possible to start from the amino acid sequence and finish with a predicted full atom structure, which is about 3–4 Å from the likely native conformation. These early attempts will doubtless be supplanted by more sophisticated realizations, but their success greatly encourages us that a solution to the protein-folding problem is indeed possible. The next few years are likely to see even more advances in this regard. In short, the future outlook for even greater progress in the protein-folding problem is rather bright.

## REFERENCES


1 Shaw, W.V., Biochem. J., 246 (1987) 1.
2 Martin, Y.C., Methods Enzymol., 203 (1991) 587.
3 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, J.R., Rodgers, J.R., Kennard, O., Simanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
4 PDB, Quarterly Newsletter, No. 61, June 1992.
5 Bowie, J.U., Luethy, R. and Eisenberg, D., Science, 253 (1991) 164.
6 Godzik, A. and Skolnick, J., Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 12098.
7 Godzik, A., Kolinski, A. and Skolnick, J., J. Mol. Biol., 227 (1992) 227.
8 Sippl, M.J. and Weitckus, S., Proteins, 13 (1992) 258.
9 Bryant, S.H. and Lawrence, C.E., Proteins, 16 (1993) 92.
10 Jones, D.T., Taylor, W.R. and Thorton, J.M., Nature, 358 (1992) 86.
11 Maiorov, V.N. and Crippen, G.M., J. Mol. Biol., 277 (1992) 876.
12 a. Skolnick, J., Kolinski, A., Brooks III, C.L., Godzik, A. and Rey, A., Curr. Biol., 3 (1993) 414.
b. Kolinski, A., Godzik, A. and Skolnick, J., J. Chem. Phys., 98 (1993) 7420.
13 Karplus, M. and Petsko, G.A., Nature, 347 (1990) 631.
14 Novotny, J., Bruccoleri, T. and Karplus, M., J. Mol. Biol., 177 (1984) 787.
15 Jernigan, R.L., Curr. Opin. Struct. Biol., 2 (1992) 248.
16 Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 219 (1991) 109.
17 Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D., J. Mol. Biol., 106 (1976) 983.
18 Wilson, C. and Doniach, S., Protein Struct. Funct. Genet., 6 (1989) 193.
19 Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 211 (1990) 959.
20 Godzik, A., Skolnick, J. and Kolinski, A., Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 2629.
21 Godzik, A., Kolinski, A. and Skolnick, J., J. Comput. Chem., 1993, in press.
22 Clementi, E., Computational Aspects for Large Molecular Systems, Lecture Notes in Chemistry, Vol. 19, Springer, Berlin, 1980.
23 Cornette, J.L., Cease, K.B., Margalit, H., Sponge, J.L., Berzowski, J.A. and DeLisi, C., J. Mol. Biol., 195 (1987) 659.

24 Warme, P.K. and Morgan, R.S.J., J. Mol. Biol., 118 (1978) 289.
25 Narayana, S.V. and Argos, P., Int. J. Pept. Protein Res., 24 (1984) 25.
26 Miyazawa, S. and Jernigan, R., Macromolecules, 18 (1985) 534.
27 Singh, J. and Thorton, J.M., J. Mol. Biol., 211 (1990) 595.
28 Rooman, M.J. and Wodak, S.J., Nature, 335 (1988) 45.
29 Bryant, S.H. and Lawrence, C.E., Proteins, 9 (1991) 108.
30 Shakhnovich, E.I. and Finkelstein, A.V., Biopolymers, 28 (1989) 1667.
31 Christensen, H. and Pain, R.H., Eur. Biophys. J., 19 (1991) 221.
32 Skolnick, J. and Kolinski, A., Annu. Rev. Phys. Chem., 40 (1989) 207.
33 Skolnick, J., Kolinski, A. and Sikorski, A., Chem. Des. Autom. News, 5 (1990) 1.
34 Skolnick, J., Kolinski, A. and Sikorski, A., Comm. Mol. Cell. Biol., 6 (1990) 223.
35 Hinds, D.A. and Levitt, M., Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 2539.
36 Skolnick, J., Kolinski, A. and Yaris, R., Biopolymers, 28 (1989) 1059.
37 Skolnick, J. and Kolinski, A., J. Mol. Biol., 221 (1991) 499.
38 Lau, K.F. and Dill, K.A., Macromolecules, 22 (1989) 3986.
39 Kendrew, J.C., Bodo, G., Dintiz, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C., Nature, 181 (1958) 662.
40 Waterman, M.S., Bull. Math. Biol., 46 (1984) 473.
41 Perutz, M.F., Kendrew, J.C. and Watson, H.C., J. Mol. Biol., 13 (1965) 669.
42 Pastore, A. and Lesk, A.M., Proteins, 8 (1990) 133.
43 Lim, V.I., Biofizika, 19 (1974) 562.
44 Chou, P.Y. and Fasman, G.D., Biochemistry, 13 (1974) 211.
45 Fasman, G.D. (Ed.) Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989.
46 Lesk, A.M. and Chothia, C., J. Mol. Biol., 136 (1980) 225.
47 Bashford, D., Chothia, C. and Lesk, A.M., J. Mol. Biol., 196 (1987) 199.
48 Lesk, A.M. and Chothia, C., J. Mol. Biol., 160 (1982) 325.
49 Chothia, C. and Lesk, A.M., J. Mol. Biol., 160 (1982) 309.
50 Gribskov, M., McLachlan, M. and Eisenberg, D.P., Proc. Natl. Acad. Sci. U.S.A., 84 (1987) 4355.
51 Altschul, S.F. and Lipman, D.J., Proc. Natl. Acad. Sci. U.S.A., 87 (1990) 5509.
52 Martinez, H.M., Nucleic Acids Res., 16 (1988) 1683.
53 Barton, G.J. and Sternberg, M.J.E., J. Mol. Biol., 198 (1987) 327.
54 Subbiah, S. and Harrison, S.C., J. Mol. Biol., 209 (1989) 539.
55 Vingron, M. and Argos, P., CABIOS, 5 (1989) 115.
56 Taylor, W.R., J. Mol. Biol., 188 (1987) 233.
57 Godzik, A. and Sander, C., Protein Eng., 2 (1989) 589.
58 Pascarella, S. and Argos, P., Protein Eng., 5 (1992) 121.
59 Argos, P., J. Mol. Biol., 197 (1987) 331.
60 Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, H.A., J. Protein Chem., 4 (1985) 23.
61 Nakai, K., Kidera, A. and Kanehisa, M., Protein Eng., 2 (1988) 93.
62 Sali, A. and Blundell, T.L., J. Mol. Biol., 212 (1990) 403.
63 Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T., Proteins, 7 (1990) 257.
64 Luethy, R., McLachlan, A.D. and Eisenberg, D., Proteins, 10 (1991) 229.
65 Ponder, J.W. and Richards, F.M., J. Mol. Biol., 193 (1987) 775.
66 Correa, P.A., Proteins, 7 (1990) 366.
67 Reid, L.S. and Thornton, J.M., Proteins, 5 (1989) 170.
68 Holm, L. and Sander, C., J. Mol. Biol., 218 (1991) 183.
69 Finkelstein, A.V. and Reva, B.A., Biofizika, 35 (1990) 402.
70 Finkelstein, A.V. and Reva, B.A., Nature, 351 (1991) 497.
71 Daopin, S., Albert, T., Baase, W.A., Wozniak, J.A. and Matthews, B.W., J. Mol. Biol., 221 (1991) 873.
72 EMBL, Protein Sequence Database, Vol. 15, EMBL Data Library, Heidelberg, Germany, 1992.
73 Needelman, S.B. and Wunsch, C.D., J. Mol. Biol., 48 (1970) 443.
74 Guss, J.M. and Freeman, H.C., J. Mol. Biol., 169 (1983) 521.

75 Baker, E.N., J. Mol. Biol., 203 (1988) 1071.
76 Adman, E.T., In Harrison, P.M. (Ed.) Metalloproteins, Part I, Macmillan, London, 1985, pp. 1–42.
77 Thornton, J.M., Flores, T.P., Jones, D.T. and Swindells, M.B., Nature, 354 (1991) 105.
78 Finkelstein, A.V. and Ptitsyn, O.B., Prog. Biophys. Mol. Biol., 50 (1987) 171.
79 Benner, S.A. and Gerloff, D., Adv. Enz. Regul., 31 (1991) 121.
80 Chirgadze, Y.N., Acta Crystallogr., A43 (1987) 405.
81 Luger, K., Szadkowski, H. and Kirschner, K., Protein Eng., 3 (1990) 249.
82 Kabsch, W. and Sander, C., Biopolymers, 22 (1983) 2577.
83 Levitt, M. and Warshel, A., Nature, 253 (1975) 694.
84 Levitt, M., J. Mol. Biol., 104 (1976) 59.
85 Hagler, A.T. and Honig, B., Proc. Natl. Acad. Sci. U.S.A., 75 (1978) 554.
86 Ycas, M., J. Protein Chem., 9 (1990) 177.
87 Go, N., Abe, H., Mizuno, H. and Taketomi, H., In Jaenicke, N. (Ed.) Protein Folding, Elsevier, Amsterdam, 1980, pp. 167–181.
88 Go, N. and Taketomi, H., Proc. Natl. Acad. Sci. U.S.A., 75 (1978) 559.
89 Abe, H., Biopolymers, 20 (1981) 1013.
90 Ueda, Y., Taketomi, H. and Go, N., Biopolymers, 17 (1978) 1531.
91 Krigbaum, W.R. and Lin, S.F., Macromolecules, 15 (1982) 1135.
92 Dashevskii, V.G., Molekulyarnaya Biologia (Translat.), 14 (1980) 105.
93 Meirovitch, H., Vasquez, M. and Scheraga, H.A., Biopolymers, 26 (1987) 651.
94 Meirovitch, H., J. Chem. Phys., 89 (1988) 2514.
95 Covell, D. and Jernigan, R.L., Biochemistry, 29 (1990) 3287.
96 Shakhnovich, E. and Gutin, A., J. Chem. Phys., 93 (1990) 5967.
97 Shakhnovich, E., Farztdinov, G. and Gutin, A.M., Phys. Rev. Lett., 67 (1991) 1665.
98 Chan, H.S. and Dill, K.A., Annu. Rev. Biophys. Biophys. Chem., 20 (1991) 447.
99 Chan, H.S. and Dill, K.A., Macromolecules, 22 (1989) 4559.
100 Kolinski, A. and Skolnick, J., J. Chem. Phys., 97 (1992) 9412.
101 Kolinski, A., Skolnick, J. and Yaris, R., J. Chem. Phys., 85 (1986) 3585.
102 Kolinski, A., Skolnick, J. and Yaris, R., Biopolymers, 26 (1987) 937.
103 Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. U.S.A., 85 (1988) 5057.
104 Sikorski, A. and Skolnick, J., Biopolymers, 28 (1989) 1097.
105 Sikorski, A. and Skolnick, J., Proc. Natl. Acad. Sci. U.S.A., 86 (1989) 2668.
106 Sikorski, A. and Skolnick, J., J. Mol. Biol., 215 (1990) 183.
107 Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. U.S.A., 86 (1989) 1229.
108 Skolnick, J. and Kolinski, A., J. Mol. Biol., 212 (1990) 787.
109 Kolinski, A., Milik, M. and Skolnick, J., J. Chem. Phys., 94 (1991) 3978.
110 Skolnick, J. and Kolinski, A., Science, 250 (1990) 1121.
111 Kuwajima, W., Proteins, 6 (1989) 87.
112 Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. and Razgulyaev, O.I., FEBS Lett., 262 (1990) 20.
113 Levinthal, C., J. Chem. Phys., 65 (1968) 44.
114 Levitt, M., Curr. Opin. Struct. Biol., 1 (1991) 224.
115 Dyson, H.J., Merutka, G., Waltho, J.P., Lerner, R.A. and Wright, P.E., J. Mol. Biol., 226 (1992) 795.
116 Matthews, C.R., In Gierasch, L.M. and King, J. (Eds.) Protein Folding, AAAS, Washington, 1990, p. 191.
117 Baumgartner, A., Annu. Rev. Phys. Chem., 35 (1984) 419.
118 Skolnick, J. and Kolinski, A., Adv. Chem. Phys., 77 (1990) 223.
119 Rey, A. and Skolnick, J., Chem. Phys., 158 (1991) 199.
120 Honeycutt, J.D. and Thirumalai, D., Proc. Natl. Acad. Sci. U.S.A., 87 (1990) 3526.
121 Binder, K., In Binder, K. (Ed.) Monte Carlo Methods in Statistical Physics, Springer, Berlin, 1986, p. 411.
122 Poland, D. and Scheraga, H.A., Theory of Helix-Coil Transitions in Biopolymers, Academic Press, New York, 1970.
123 Handel, T. and DeGrado, W.F., Biophys. J., 61 (1992) A265.
124 Raleigh, D.P. and DeGrado, W.F., J. Am. Chem. Soc., 114 (1992) 10079.
125 Rey, A. and Skolnick, J., J. Comput. Chem., 13 (1992) 443.

126 Hecht, M.H., Richardson, J.S., Richardson, D.C. and Ogden, R.C., Science, 249 (1990) 884.
127 Sander, C. (Ed.) Protein Design Exercises, Vol. 1, EMBL, Heidelberg, 1986.
128 Covell, D.G. and Jernigan, R.L., Biochemistry, 29 (1990) 3287.
129 Milik, M. and Skolnick, J., Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 9391.