J. Mol. Biol. (1992) 227, 227–238

Topology Fingerprint Approach to the Inverse Protein Folding Problem

Adam Godzik[1], Andrzej Kolinski[1,2] and Jeffrey Skolnick[1]†

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

[2]*Department of Chemistry, University of Warsaw Pasteura 1, 02-093, Warsaw, Poland*

(Received 17 March 1992; accepted 28 April 1992)

We describe the most general solution to date of the problem of matching globular protein sequences to the appropriate three-dimensional structure. The screening template, against which sequences are tested, is provided by a protein "structural fingerprint" library based on the contact map and the buried/exposed pattern of residues. Then, a lattice Monte Carlo algorithm validates or dismisses the stability of the proposed fold. Examples of known structural similarities between proteins having weakly or unrelated sequences such as the globins and phycocyanins, the eight-member α/β fold of triose phosphate isomerase and even a close structural equivalence between azurin and immunoglobulins are found.

Keywords: protein structure prediction; protein stability; plastocyanin-azurin-immunoglobulin similarity; globin-phycocyanin similarity; TIM barrel similarity

1. Introduction

Most successful attempts at predicting the structure of a newly sequenced protein try to identify another protein, with known three-dimensional structure, which is related to the protein of interest. Such sequence homology techniques have been very successful at detecting even weak similarities between related sequences (Taylor, 1986; Argos, 1987; Gribskov *et al.*, 1987; Altschul & Lipman, 1990). Extension of this approach to detect structural similarities between proteins having little or no sequence homology requires techniques that incorporate tertiary structural information (Thornton *et al.*, 1991). We present here a coupled methodology, which achieves this objective, and address another important and related problem in homology modeling, that of verifying the stability of the proposed topology.

In schematic outline, this method proceeds as follows: first, one aligns the given sequence with the framework of a known protein structure. This process is then repeated for a library of structures, and the best tentative assignment is made based on the estimated energy of each of them. Then, a lattice model of the three-dimensional structure of the proposed alignment is built, and the stability of the model is investigated by dynamic Monte Carlo simulations. Because the method is based on the interaction pattern as defined by the contact maps, it can match proteins having no apparent sequence similarity. Examples of correctly identified structural similarities include the phycocyanin/globin similarity, a number of α/β proteins having the TIM fold, and the relationship of the azurin, plastocyanin and immunoglobulin folds.

Previously, Ponder & Richards (1987) have developed a tertiary template method which analyzes the packing of different protein sequences within a given protein core. While very powerful in principle, it is prohibitively expensive and focuses only on a single feature of protein structure. Another method (Finkelstein & Reva, 1991), matches protein sequences to an idealized protein fold. Given that the rules that define a given topology are not well understood, it is perhaps safer to use real proteins rather than idealized ones in the assessment of the interaction pattern. This would also permit the more accurate construction of the protein structure should the technique prove successful. Bowie *et al.* (1991) use a template protein to calculate a table of three-dimensional/one-dimensional propensities that reflect the probability of a given amino acid being in a particular environment, defined in terms of solvent accessibility and secondary structure. Dynamic programming techniques then match the sequence to the pattern of

† Author to whom all correspondence should be addressed.

0022-2836/92/170227-12 $08.00/0

environments defining the fold. However, because these environments are static and reflect the template protein, they may be substantially different in the actual fold of the protein of interest. Indeed, as shown below, there are examples where this approximation in the context of our methodology leads to incomplete structural alignments. Finally, and common to all these methods, is the problem of assessing the stability of a conjectured fold. The ideal methodology should provide some confidence that the predicted fold is really stable and not apparently stable, simply because it is the least worst in the library of known template structures.

2. Methods

(a) *Structural fingerprint algorithm*

The 3-dimensional structure of any protein defines a specific interaction pattern among the residues which, in turn, can be described by a 2-dimensional map, often used in the literature under different names, such as distance plots or matrices, or contact maps. These maps are used to stress the similarity between 2 different structures or to study changes in structure due to ligand binding (Phillips, 1970; Nikishawa *et al.*, 1972; Richards & Kundrot, 1988; Godzik & Sander, 1989; Scharf, 1989). In the present contribution, we use contact maps, where the interactions (contacts) between amino acid side-chains are recorded as binary information (2 residues are or are not interacting). We adopt the definition that 2 side-chains are in contact whenever any heavy atom in the 1st side-chain is closer than 5 Å (1 Å = 0·1 nm) to any heavy atom in the 2nd. Provided that one has a reasonable interaction energy scale, as described in further detail below, the contact map can also be used to estimate the total energy of any sequence folding into any given structure. If this holds, then one can then test the conjecture that the contact map possesses sufficient intrinsic information about the protein topology so that unrelated sequences having similar 3-dimensional structures can also be recognized.

In this spirit, we approximate the total energy of a folded protein as the sum of contributions arising from: (1) burying hydrophobic residues and exposing polar ones to the solvent (this pattern is related to, but is not identical with, the specification of the secondary structure), (2) pairwise interactions between all contacting side-chain residues, (3) clusters of interacting triplets of residues. Surprisingly for some triplets, the contribution of the 3-body term is non-negligible. For example, some buried, oppositely charged side-chains would rather be exposed to solvent, but given that they are buried, they prefer to interact with some particular hydrophobic groups.

Schematically, the energy of protein A is given by:

$$E = \sum_i \Gamma_i^{\mathrm{A}} E_1(\mathrm{A}_i) + \sum_i \sum_{>j} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} E_2(\mathrm{A}_i, \mathrm{A}_j) + \sum_i \sum_{>j} \sum_{>k} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} \mathbf{C}_{\mathbf{ik}}^{\mathbf{A}} \mathbf{C}_{\mathbf{kj}}^{\mathbf{A}} E_3(\mathrm{A}_i, \mathrm{A}_j, \mathrm{A}_k), \quad (1)$$

where **i**, **j** and **k** are numbers of positions along the sequence; A_i, A_j and A_k are amino acids found at these respective positions in protein A; $\Gamma_{\mathbf{i}}^{\mathbf{A}}$ is the buried/exposed classification of position **i** in protein A, and $\mathbf{C}_{\mathbf{ij}}^{\mathbf{A}}$ is the binary information about the contact between positions **i** and **j** in protein A. Finally, E_1, E_2 and E_3 are, respectively, the 1, 2 and 3 body contributions to the total energy; a listing of these parameters may be obtained from the authors by anonymous FTP.

The set of energetic parameters are derived from a statistical analysis of a structural database comprised of 59 non-homologous proteins (Hobohm *et al.*, 1992) with all protein structures extracted from the Protein Data Bank (Bernstein *et al.*, 1977). Typically, the 1, 2 and 3 body terms contribute 40%, 20% and 40%, respectively, to the total energy. The relative contributions emerge from the statistics of the contacts themselves, and there are a number of instances where the 3 body terms aid in identifying the correct fold. A detailed dissection of the importance of each term to the overall stability is the subject of another paper (Godzik & Skolnick, 1992). Originally, this energy scale was developed for lattice folding simulations (Skolnick & Kolinski, 1989), but it has proven to be very efficient in identifying sequences that are likely to fold to a similar structure.

The evaluation of eqn (1) requires knowledge of the buried and exposed residues and a list of all interacting pairs and triplets. This information is calculated for every structure in the database and is stored as a "structure" or "topology fingerprint" library. This fingerprint does not use sequence information, but merely defines the characteristics of each position along the template chain. Thus, one can calculate the energy of a corresponding system (protein B) having the same interaction pattern, but a different sequence. For this situation, eqn (1) becomes:

$$E = \sum_i \Gamma_i^{\mathrm{A}} E_1(\mathrm{B}_i) + \sum_i \sum_{>j} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} E_2(\mathrm{B}_i, \mathrm{B}_j) + \sum_i \sum_{>j} \sum_{>k} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} \mathbf{C}_{\mathbf{ik}}^{\mathbf{A}} \mathbf{C}_{\mathbf{kj}}^{\mathbf{A}} E_3(\mathrm{B}_i, \mathrm{B}_j, \mathrm{B}_k). \quad (2)$$

where B_i, B_j and B_k are now the amino acids found at the respective positions in the sequence of protein B, but the $\Gamma_{\mathbf{i}}^{\mathbf{A}}$ and $\mathbf{C}_{\mathbf{ij}}^{\mathbf{A}}$ are taken from the structure of protein A. On the basis of the lowest energy, this procedure matched each sequence with its corresponding structure in the fingerprint library (Godzik & Skolnick, 1992).

We then applied the method to a larger database of 125 proteins, of which 22 were in the original database. Again, on the basis of the lowest energy, all sequences were correctly matched to their native structure. Thus, the minimal test of the ability of a sequence to recognize its own fold was passed. However, without the possibility of gaps, the method fails to recognize related folds having insertions and deletions. For example, it does not identify the structural relationship between azurin and plastocyanin (Guss & Freeman, 1983), which have identical topology, with the former having a helix hairpin insertion (Chothia, 1982).

Gaps and insertions can be accommodated by omitting certain residues from the energy calculation (or by allowing for inert residues). As in standard sequence homology techniques (Needleman & Wünsch, 1970), different penalties are imposed for introducing a gap/insertion and for its extension. However, the 2nd and 3rd terms in eqn (1) depend on an alignment that occurs further on and/or earlier in the sequence. Thus, we introduce the "frozen" approximation, with the energy calculated according to eqn (3):

$$E = \sum_i \Gamma_i^{\mathrm{A}} E_1(\mathrm{B}_i) + \sum_i \sum_{>j} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} E_2(\mathrm{B}_i, \mathrm{A}_j) + \sum_i \sum_{>j} \sum_{>k} \mathbf{C}_{\mathbf{ij}}^{\mathbf{A}} \mathbf{C}_{\mathbf{ik}}^{\mathbf{A}} \mathbf{C}_{\mathbf{kj}}^{\mathbf{A}} E_3(\mathrm{B}_i, \mathrm{A}_j, \mathrm{A}_k), \quad (3)$$

with the notation as before. Here, the energies of the amino acids from sequence B are calculated as if they would interact with their partners from protein A. This

approximation may be rationalized by arguing that in similar proteins, the environment of 2 equivalent positions should on average tend to be similar. Now, the energy can be broken down to separate contributions coming from each amino acid:

$$E(B_i) = \Gamma_i^A E_1(B_i) + \sum_j C_{ij}^A E_2(B_i, A_j) + \sum_j \sum_{>k} C_{ij}^A C_{ik}^A C_{kj}^A E_3(B_i, A_j, A_k), \quad (4)$$

where the summations are performed only for the subset of interacting pairs (triplets) for which **i** is one of the partners, and care must be taken not to overcount pair and triplet contributions when calculating the total energy. Eqn (4) can be interpreted as defining the score of aligning amino acid B_i with position **i** in protein A, and it is a form of position dependent amino acid similarity table, as has been used in profile methods (Gribskov *et al.*, 1987; Bowie *et al.*, 1991). Following the original Needleman & Wünsch (1970) approach, a dynamic programming algorithm was developed to search for optimal alignment between an arbitrary sequence and a given topology fingerprint.

As they have been obtained within the context of the frozen approximation, these alignments represent the best fit to the static environment of the original protein A. Of course, in the real protein, the actual environment in protein B may differ from that in protein A. There are several possibilities of going beyond this restrictive frozen approximation. One of them, currently implemented in our program, first requires a tentative alignment of A and B obtained, for instance, with the help of the frozen approximation:

$$\begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & - & - & A_7 & A_8 & A_9 \\ B_1 & B_2 & B_3 & - & B_4 & B_5 & B_6 & B_7 & B_8 & - & B_9 \end{array}$$

We shall describe this alignment by the function $N_{A\to B}(i) = j$, which maps residue i of sequence A to residue j in sequence B. So, for instance, in the example above, $N_{A\to B}(5) = 4$ and $N_{A\to B}(7) = 8$. Thus, the score for each position, as calculated by eqn (4), is updated to:

$$E(B_i) = \Gamma_i^A E_1(B_i) + \sum_j C_{ij}^A E_2(B_i, B_{N_{A\to B}(j)}) + \sum_j \sum_{>k} C_{ij}^A C_{ik}^A C_{kj}^A E_3(B_i, B_{N_{A\to B}(j)}, B_{N_{A\to B}(k)}), \quad (5)$$

which represents the interaction of each residue within a new, updated environment. A new and (hopefully) better alignment is now obtained with scores calculated according to eqn (5), a new function $N_{A\to B}$ is calculated, and the procedure can be repeated. This procedure usually converges within 5 to 10 iterations, and for similar sequences usually does not involve large shifts. However, for some examples discussed later in the paper, the alignment can be substantially modified.

A schematic overview of the whole topology fingerprint matching methodology is presented in Fig. 1. For mixing and matching of sequences to their structure in the fingerprint library, again we are able to assign all protein sequences to their correct fold both for the small database used for energy parameter development and for the larger database. As they should be, the resulting alignments are devoid of gaps. This approximation renders this stage of the algorithm logically equivalent to that of Bowie *et al.* (1991). However, we employ a different definition of the residue environment, which permits further analysis after the interaction parameters are "thawed", i.e. the pairs and triplets partners are replaced by the actual residues in the test sequence.

Note, however, that the buried/exposed pattern in the 2nd pass remains invariant, while the partners in the 2 and 3 body terms reflect those actually present in the test sequence. Presumably, if the template and test protein have similar structures, the packing interactions in the interior of the protein change the most, and the buried/exposed pattern changes the least. However, since we subsequently perform Monte Carlo refinements on the contact map or lattice Monte Carlo refinements (see below), the buried/exposed pattern changes as well. Alternatively, one can use a contact number based energy scale where residues having a number of contacts below (exceeding) a threshold value are classified as exposed (buried). Then in the 2nd pass, depending on the identity of the residue and the number of contacts it experiences, the buried/exposed pattern also changes.

Our alignment program was extended to permit the rapid scanning of a large sequence database (Protein Sequence Databank, 1991). The final score is the normalized energy difference between the best alignment of the sequence to the template and the same sequence after multiple randomizations (Waterman, 1984). For simplicity, the gap penalties of loops and secondary structures are identical, and only those alignments involving the entire template are considered. The best results are obtained after averaging the score for different gap penalties.

The method for identifying a likely fold proceeds as follows: A given sequence is tested against the entire fingerprint library of the 125 fingerprint proteins, and the best scoring structure is selected. Then, the best fitting fingerprint was run against the entire sequence database. We note that running a sequence against a given fingerprint typically takes 20 s of CPU time on a Sparc2 workstation. If the sequence score lies in the tail and not in the Gaussian portion of the energy distribution (see Fig. 2, for example), this indicates the sequence should be further examined. The screening of a structure against the entire sequence database also provides possible sequences that are compatible with a given fold. We have also examined the results of screening a given structure against the entire sequence data base. Those sequences having the top scores were subjected to further analysis. For both cases, simple iterations where, after the 1st pass alignment, the actual partners are employed in eqn (1) for the next alignment iteration, as well as Monte Carlo minimizations, are used to find the best alignment.

(b) *Lattice refinement*

The alignment of the test protein sequence to the structural fingerprint provides the starting point for the next procedure, which circumvents many of the limitations inherent in the 1st step (or in any quasi-1-dimensional scheme). The structural fingerprint provides a template onto which the 3-dimensional structure of the protein projected onto the lattice is built and subsequently refined (Skolnick & Kolinski, 1990). We currently use a hybrid lattice that is intrinsically capable of representing carbon backbones at the level of 1 Å root-mean-square (r.m.s.†) (Skolnick & Kolinski, 1991). While our earlier lower resolution lattice models used local propensities for secondary structure that favored native-like conformations, here we only employ predicted information about the relationship between the test

† Abbreviations used: r.m.s., root-mean-square; Ig, immunoglobulin; MC, Monte Carlo.

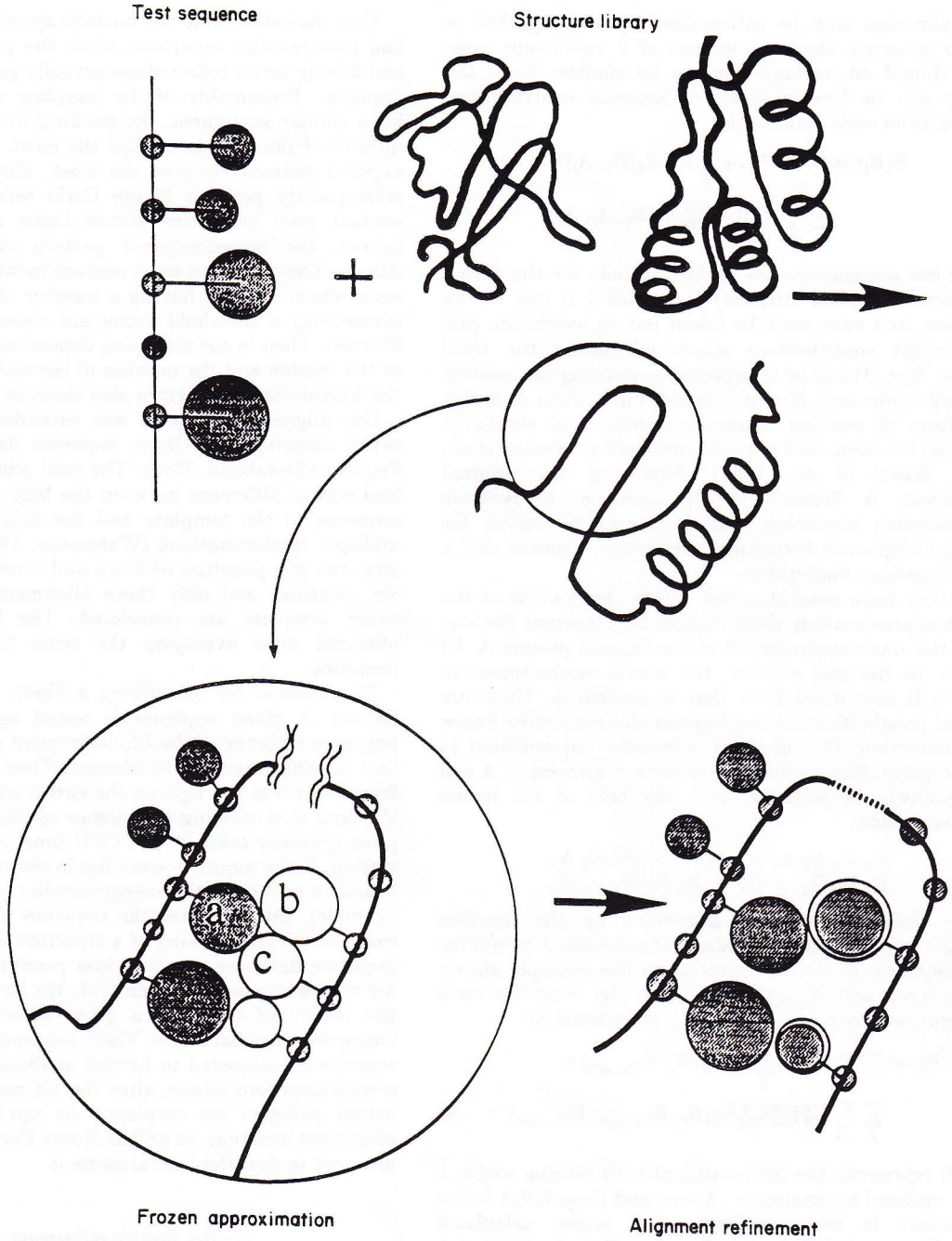


Figure 1. Schematic representation of the template matching procedure. A test sequence is threaded through the structural fingerprint library derived from 3-dimensional protein structures. Initially, the frozen approximation is used in which side-chains in the test sequence (shaded circles) interact with their partners (open circles) *b* and *c* of the original template structure. The alignment may have gaps. The best alignments based on an energy criterion are then subjected to further analysis where all the interacting residues are replaced by the actual partners. Thus, for example, a *c* type 3 body interaction in the frozen approximation does not occur in the actual refined alignment.

sequence and template protein. Thus, the last vestiges of a "built in" answer are removed, and what is presented below are truly blind predictions.

As shown in Fig. 2(a), the protein is embedded into an underlying cubic lattice where adjacent points are connected by a vector of the type (1, 0, 0), and the centers of consecutive α-carbons (open circles) are joined by vectors of the type $(\pm 2, \pm 1, 0)$, $(\pm 2, \pm 1, \pm 1)$, or $(\pm 1, \pm 1, \pm 1)$. Thus, there are in principle 56 possibilities for every backbone vector. If realistic distance constraints reflecting the distribution of 2nd nearest neighbor α-carbon distances (r_{13}) are used, then there are about 30 intrinsic rotational states per bond. The underlying distribution of r_{13} reflecting the degeneracy of the lattice states is quite close to that seen in protein crystal structures and requires minor modifications to reproduce this distribution. Of course, this is now a variable bond length model, but reasonable correspondence to real proteins is achieved by equating the average virtual bond length (close to a (2, 1, 0) vector) to the C^{α}–C^{α} virtual bond length of 3·785 Å. In this case, the spacing between cubic lattice points equals 1·7 Å. In addition to the central lattice site,

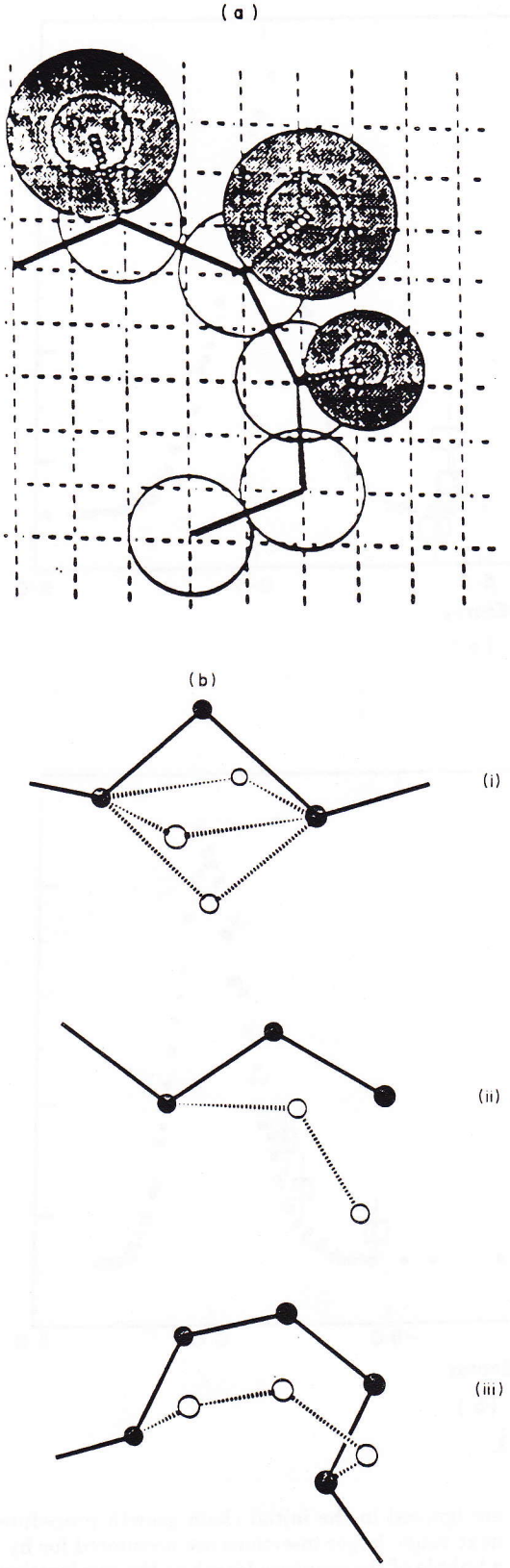


Figure 2. (a) Lattice representation of a fragment of a 3-dimensional protein structure. The α-carbon virtual bonds belong to the set of basis vectors (2, 1, 0), (2, 1, 1) and (1, 1, 1) embedded onto the underlying cubic lattice. The open circles represent the hard core envelope of the main-chain, and the filled circles represent the side-groups. The outer (inner) circle denotes the attractive (repulsive) region. The size depends on the amino acid type, and the orientation is coupled to the main-chain conformation. Multiple rotamers are permitted. (b) The allowed local conformational rearrangements of the protein backbone employed in the Monte Carlo procedure, consisting of (i) spike moves, (ii) end moves and (iii) 4 bond moves. For the sake of clarity, the side-chains are not displayed.

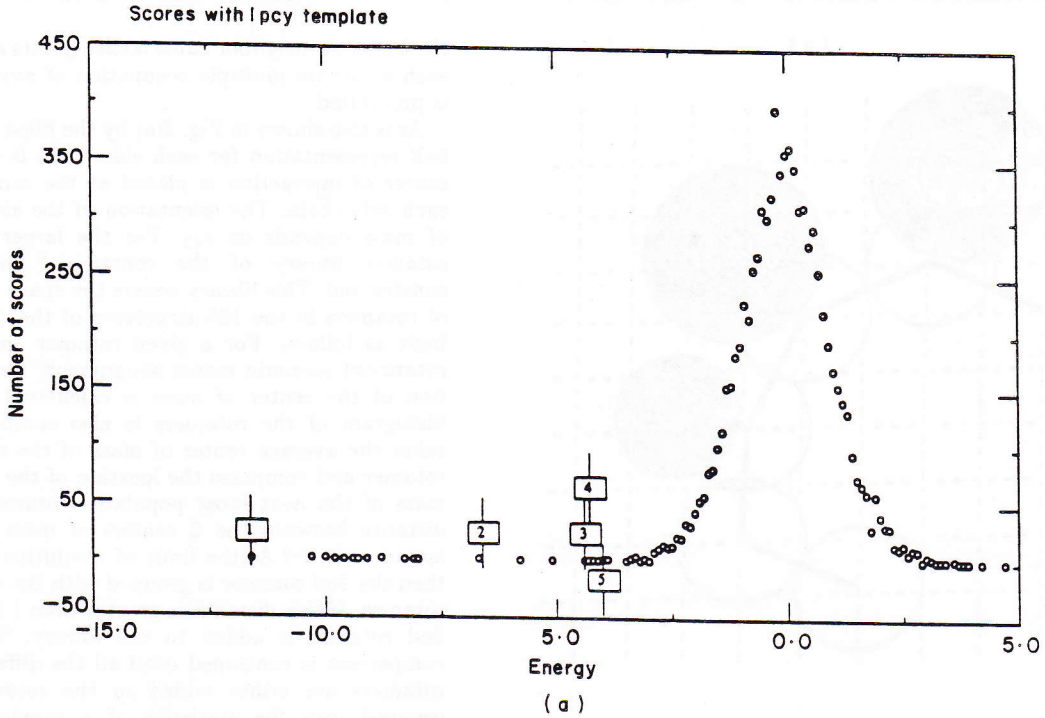
the 6 nearest-neighbor cubic lattice points are occupied by each α-carbon; multiple occupation of any backbone site is prohibited.

As is also shown in Fig. 2(a) by the filled circles, a single ball representation for each side-chain is employed. The center of interaction is placed at the center of mass of each side-chain. The orientation of the side-chain center of mass depends on r_{13}. For the larger side-chains, a rotamer library of the centers of mass has been constructed. This library covers the space of all examples of rotamers in the 125 structures of the database and is built as follows: For a given rotamer and based on its rotational isomeric states assignment, the average location of the center of mass is calculated. A population histogram of the rotamers is also compiled. One then takes the average center of mass of the most populated rotamer and compares the location of the mean center of mass of the next most populated rotamer to it. If the distance between the 2 centers of mass is less than 1 lattice unit, 1·7 Å (the limit of resolution of the model), then the 2nd rotamer is grouped with the most populated rotamer. If this distance is greater than 1 lattice unit, the 2nd rotamer is added to the library. This process of comparison is continued until all the different side-chain rotamers are either added to the rotamer library or grouped into the statistics of a previously examined rotamer. These centers of mass are then rotated into the appropriate co-ordinate system of all the various lattice states and stored as a table.

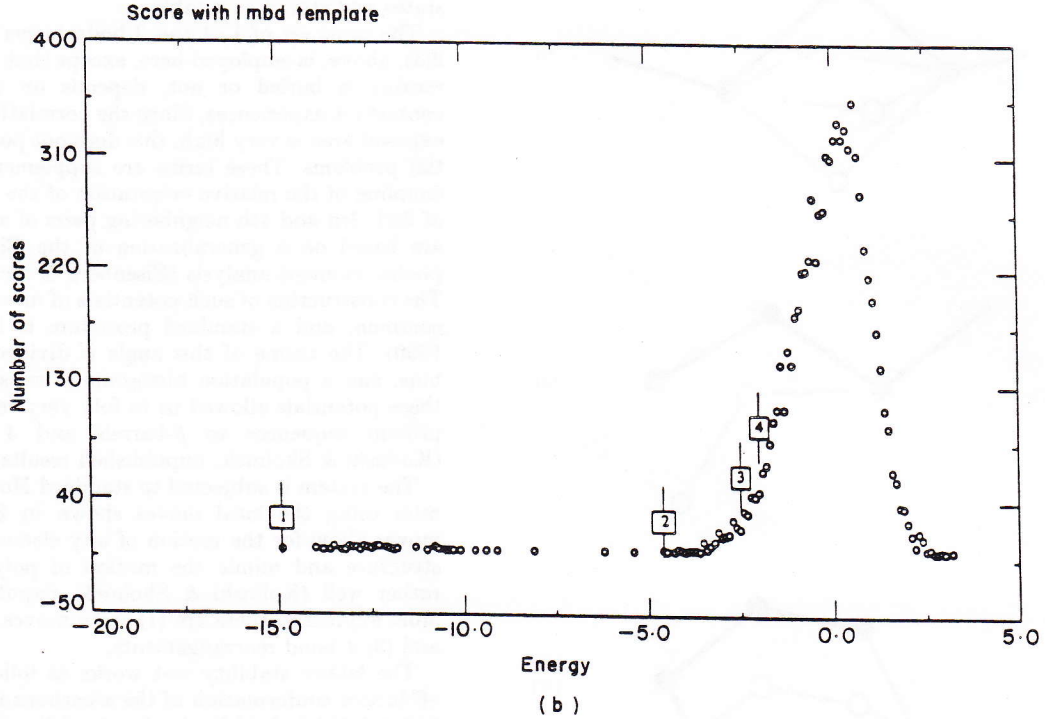
The same set of 1, 2 and 3 body terms used in section 2(a), above, is employed here, except that now, whether a residue is buried or not, depends on the number of contacts it experiences. Since the correlation with buried/exposed area is very high, this does not pose any substantial problems. These terms are supplemented by a local coupling of the relative orientation of the centers of mass of 2nd, 3rd and 4th neighboring pairs of amino acids and are based on a generalization of the Eisenberg hydrophobic moment analysis (Eisenberg & McLachlan, 1986). The construction of such potentials of mean force is rather common, and a standard procedure is followed (Sippl, 1990). The cosine of this angle is divided into 10 equal bins, and a population histogram is constructed. Use of these potentials allowed us to fold very regular, idealized protein sequences to β-barrels and 4 helix bundles (Kolinski & Skolnick, unpublished results).

The system is subjected to standard Monte Carlo dynamics using the local moves shown in Fig. 2(b). These moves allow for the motion of any element of secondary structure and mimic the motion of polypeptide chains rather well (Kolinski & Skolnick, unpublished results). More explicitly, there are (1) spike moves, (2) end moves, and (3) 4 bond rearrangements.

The lattice stability test works as follows: The initial off-lattice conformation of the α-carbons is built onto the 210/211/111 hybrid lattice by the following procedure in which the excluded volume of the backbone is rigorously maintained: The PDB file of the template, protein A, α-carbon co-ordinates are converted into lattice units. Then, the equivalence mapping of residues in B to the



(a)



(b)

Fig. 3.

template protein, A. $N_{A\to B}$ is used as a set of co-ordinate constraints for the construction of the initial α-carbon conformation on lattice. Chain construction begins at the N terminus for the 1st set of equivalent residues, and the chain is sequentially grown. If the jth residue has a partner i, then it is placed at the closest lattice position subject to excluded volume constraints. Gaps have to be treated separately. If the gap is a single residue long, e.g. j-2 and j have partners in A, then residue j and j-1 are simultaneously constructed. If not, the inserted residues are ignored in the initial chain growth procedure. In the next stage, larger insertions are accounted for by growing a bubble of the requisite length at the gap location. This is readily achieved using 4 bond moves, where a set of 3 bonds is replaced by a set of 4 bonds; the process is iterated until the entire gap is spanned. This α-carbon trace is then supplemented by side-chains, and the system is allowed to relax using the Monte Carlo dynamics embodied in Fig. 2(b). Long runs are performed to determine whether or not the initial conformation is stable.

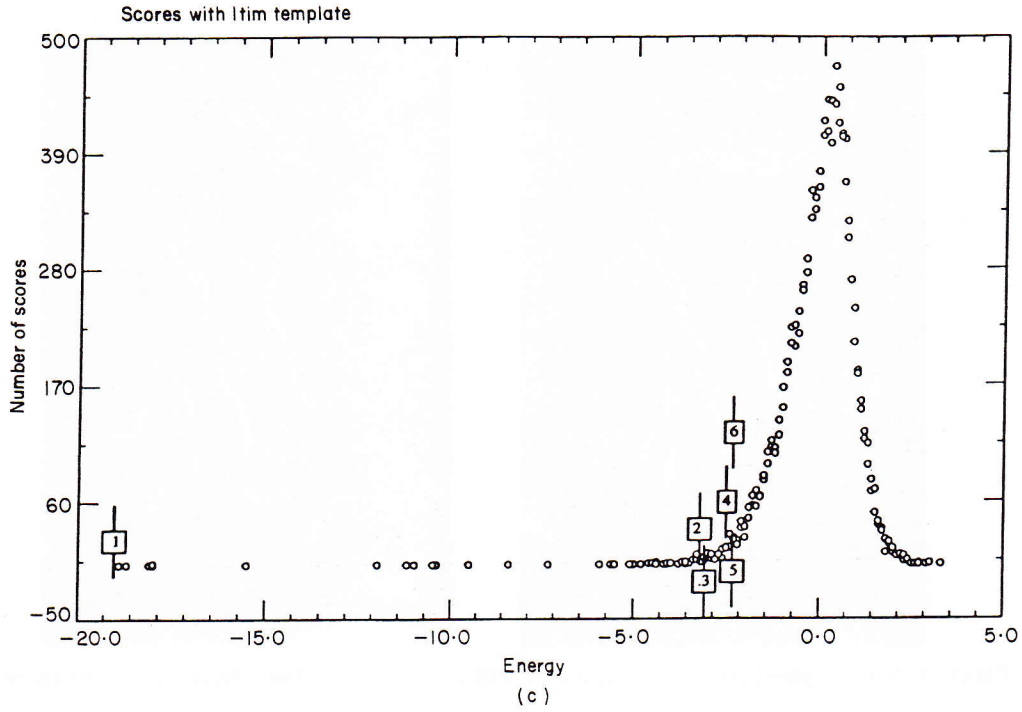


Figure 3. Plot of the number of scores in (a), (b), (c) using 1pcy, 1mbd and 1tim, respectively, as the structural fingerprint *versus* the number of standard deviations from the mean energy after randomizations. The highest scores for different sequence families are indicated in the Figure. (a) 1, plastocyanin; 2, amicyanin; 3, azurin; 4, pseudoazurin; 5, Ig kappa. (b) 1, myoglobin; 2, hemoglobin; 3, phycocyanin; 4, leghemoglobin; (c) 1, triose phosphate isomerase; 2, ribulose-1,5-biphosphate oxygenase; 3, indole synthase; 4, enolase; 5, alpha-amylase; 6, tryptophan synthase.

3. Results

(a) *Plastocyanin fingerprint and related studies*

Using the X-ray structure of poplar plastocyanin (Guss & Freeman, 1983) as the fingerprint, the entire methodology was employed to investigate which sequences are compatible with the plastocyanin fold. There are a number of known plastocyanin structures, and several other proteins have this topology. For example, azurin exhibits the same eight member, Greek key, β-barrel fold with a large insertion (Chothia, 1983); yet, the sequence similarity is very weak and can only be recovered *via* multiple sequence alignments of the whole families of plastocyanins and azurins.

Figure 3(a) shows the results of the initial screening of all sequences in the sequence database against the plastocyanin fingerprint. The screening identified all (Chothia, 1982) plastocyanin sequences in the database, followed by all amicyanins (Van Beeumen *et al.*, 1991), H.8 outer membrane precursor proteins (with the correctly identified azurin-like domain) (Kawula *et al.*, 1987), azurins (Baker, 1988) and pseudoazurins (Adman *et al.*, 1989). Thus, all the top scoring sequences have the same core topology as plastocyanin (Bairoch, 1989). The first non-obvious match was the Ig kappa chain and, in fact, the whole immunoglobulin family scored consistently high in all comparisons (see below). The only known homolog of plastocyanin which was mixed into uncertain alignments was cytochrome *c* oxidase (Holm *et al.*, 1987). We then examined two proteins of known structure that scored well but which do not have the plastocyanin topology: dihydrofolate reductase and protease B. In fact, these proteins score high, not because of global alignments of their topology, but due to short, very good alignments of β hairpins. In the former case, the C-terminal hairpin from residue 132 to 159 is identified. For protease B, two hairpins involving residues 16 to 46 and 46 to 56 appear. This suggests another, and possibly more important, use of the alignment method as a super-secondary structure predictor; this promising application will be further explored.

The second level MC/iteration alignment was applied to the azurin sequence and the plastocyanin fingerprint. The alignment equivalenced 96 positions, which translates into an r.m.s. between these positions of 6·0 Å. This is a superset of the best r.m.s. structural alignment, but the current scheme cannot identify the core residues. While the first pass worked rather well for plastocyanin analogs, in the case of papain–cathepsin matching (Musil *et al.*, 1991), the first pass only equivalences the C termini, whereas on the second pass, all the appropriate regions including the N termini are equivalenced, with the energy dropping from −13·7 to −40·3 kT.

The predicted plastocyanin–azurin alignment was then employed to build the three-dimensional lattice model of the azurin structure. The 30 inserted residues in azurin do not have any equivalent positions in plastocyanin and were initially assigned random positions subject to connectivity

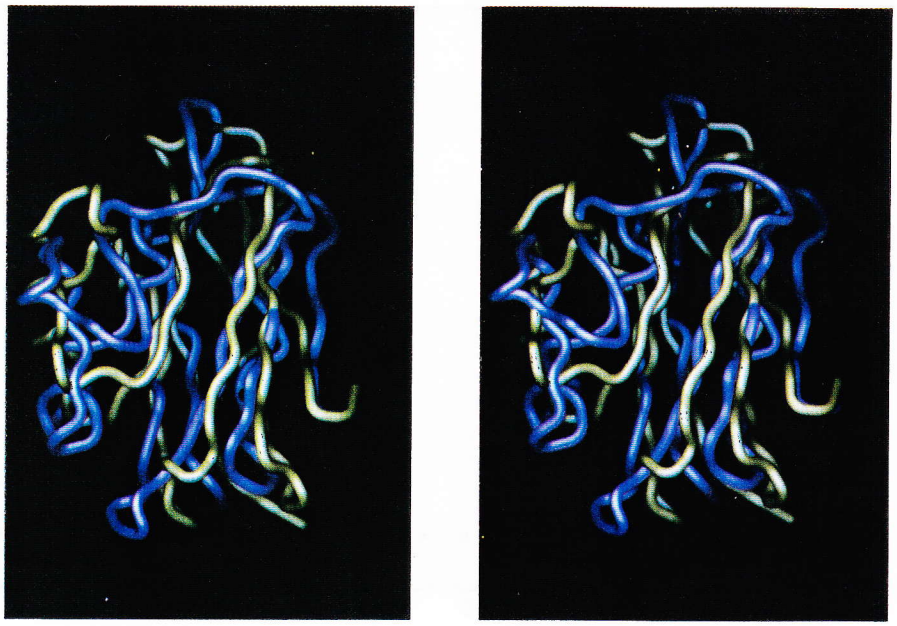


Figure 4. Superimposed predicted azurin structure (yellow) and actual crystal structures (magenta).

and excluded volume constraints. During the course of the Monte Carlo (MC) simulation, the whole structure rearranged somewhat, with more than 5·8 Å r.m.s. between the starting and final configurations, but with no changes in topology. After several long MC runs, the r.m.s. difference between the model and the known crystal structure of azurin decreased from 8·3 to 6·2 Å; a superposition of the model onto the native structure is shown in Figure 4. These values have to be compared to the inherent resolution of the model, which can be checked by performing long MC simulations of plastocyanin starting from the correct crystal structure. For plastocyanin itself, after very long simulations, the structure remained stable with the same topology, and the r.m.s. fluctuates around 6 Å from native, which is mostly due to the lack of sufficient specificity in the tertiary interaction scale.

An important question is whether the current class of lattice models can distinguish correctly from incorrectly predicted topologies. Thus, the same experiment was performed on a spurious (in the sense that conformation is incorrect) but relatively high scoring alignment (in the top 800) of plastocyanin onto the monomer of superoxide dismutase, another Greek key β-barrel protein (Getzoff *et al.*, 1989) with different β-sheet packing. As previously, the starting conformation is derived on the basis of the alignment procedure. However, in contrast to the plastocyanin sequence having the correct conformation, this structure slowly dissolved, and diverged from the initial conformation by more than 12·7 Å. The conformation started at 15·4 Å and finished at 14·1 Å from the native state. Thus, while the tertiary interaction scheme needs improvement, nevertheless, it can recognize correct from incorrect topologies of the given sequence.

One initially concerning point was the high fidelity with which the immunoglobulin chains appeared to align with plastocyanin. Since the structure of the Ig lambda chain (1RHE) is known (Furey *et al.*, 1983), this prediction of structural similarity can be checked, and has been confirmed. Thus, while it was known that the two were in the same antiparallel β-family (Lesk, 1991), their close structural similarity had not been noticed. For distantly related sequences, it is not at all clear what is the best criterion for structural superposition; as shown below, different criteria provide different alignments; this point will be elaborated on elsewhere (Godzik & Skolnick, unpublished results). However, all the gross features such as secondary structural elements coincide. The alignment obtained from the fingerprinting method is given below with the plastocyanin (2RHE) sequence on the top (bottom). The procedure renders 87 residues equivalent, and identifies 70 correct side-chain contacts for an overall r.m.s. of 7·01 Å.

```
IDVLLGADDGSLAFVP SEFSISPGEKIVFKNN     AGF PHNIV  F  DEDSIPSGVDASK       1pcy
          .. ... ...| .||... ..        .|. . ...  .  . ....  . ...
        ESVLTQPPSASGTPGQRVTISCTGSATDIGSNSVIWYQQVPGKAPKLL  IYYND       2rhe

ISMSE EDLL  NAKGE  TFEVA LSNKGEYSFYCSPHQ    GAGMVG  KV TVN            1pcy
.. |  .|..  ... .  .. .. |... |...|| ..      .|. |  |. ..
LLPSGVSDRFSASKSGTSASLAISGLESEDEADYYCAAWNDSLDEPGFGGGTKLTVLGQPK         2rhe
```

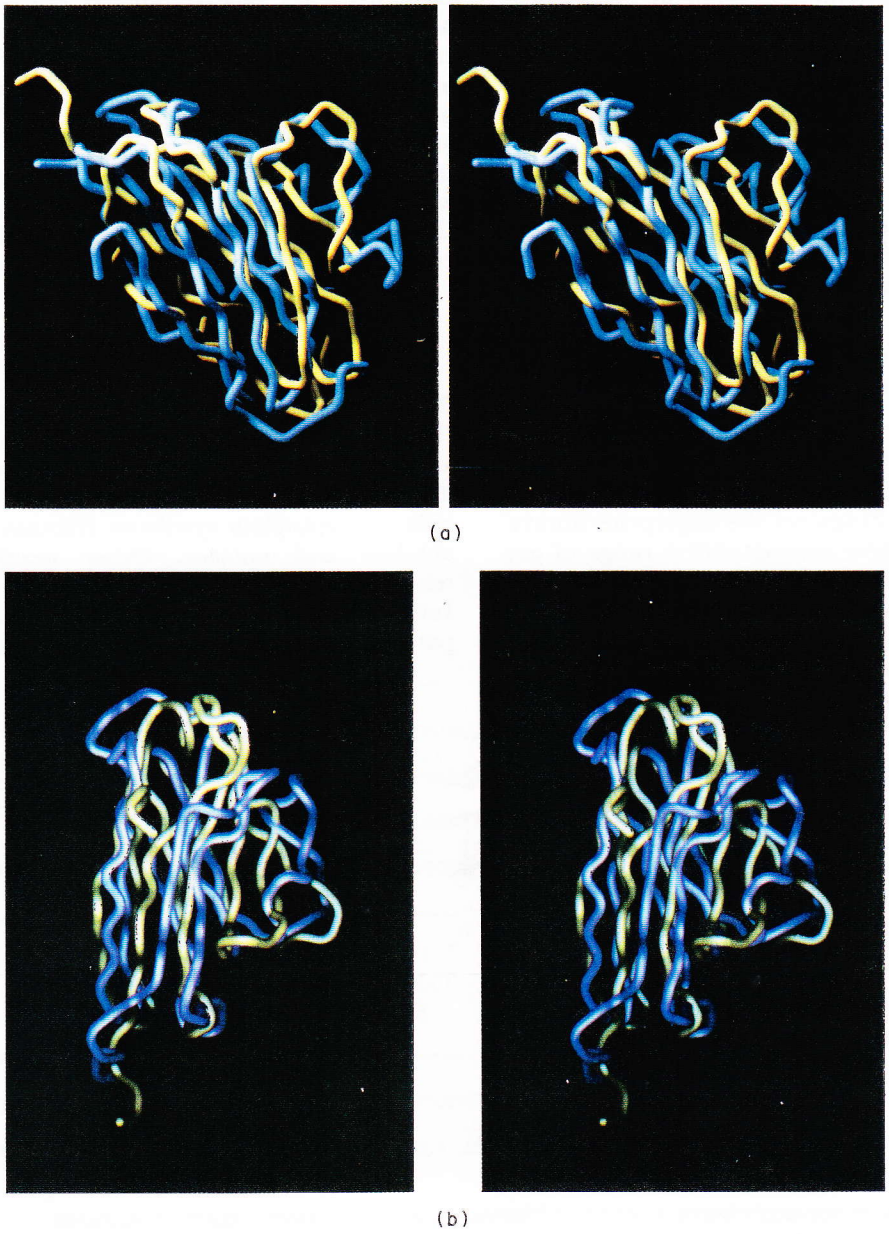


Figure 5. (a) Superimposed crystal structures of azurin (blue) and 1RHE (yellow). (b) Superimposed structures of predicted (magenta) and actual crystal (yellow) structures of 1RHE.

Note that direct superposition of the two structures on the basis of minimizing the overall r.m.s. gives an r.m.s. between them of 2·99 Å with 44 overlapping contacts. Alignment of the contact maps relates 76 pairs of residues with an overall r.m.s. of 6·72 Å and 140 superimposed contacts.

Furthermore, the azurin–immunoglobulin alignment is so good (with the N-terminal β-strand of plastocyanin is almost deleted, and there is a β-hairpin insertion where the helix-loop insertion of azurin in plastocyanin occurs) that we attempted to model the immunoglobulin structure based on azurin and proceeded to use azurin as the structural fingerprint. The resulting alignment of the two molecules is:

```
AQCEATIESNDAMQY  DLKEMVVDKSCKQFTVHLKHVGKMAKSAMGHNWVLTKEADKEGVATDGMNAGLAQ      2aza
                 ..  ... .. ..  ... .... .   | ..    ..... . .  |...
               ESVLTQPPSASGTP GQRVTISCTGS   ATDIGSNSVIWYQQVPGK  APKLL          2rhe

DYVKAGDTRVI    AHTKVIGGGESDSVTFDVSKLTPGEAYAYFCSFPGHWAMMKGTL    KLSN          2aza
.|... ..  .   .. .     .  |... .| |.. ... |.|. .    ..   .    ||
IYYNDLLPSGVSDRFSAS    KSGTSASLAISGLESEDEADYYCAAW NDSLDEPGFGGGTKLTVLGQPK      2rhe
```

The procedure aligns 100 of the 114 residues in 1RHE for an r.m.s. of 6·71 Å. The superposition of the two crystal structures is given in Figure 5(a). Furthermore, the 1RHE conformation after lattice Monte Carlo refinement is 5·4 Å from the immunoglobulin native structure; the superposition is shown in Figure 5(b). Note that superimposing azurin and immunoglobulin based on an r.m.s. criterion gives an r.m.s. of the two structures of 3·74 Å with only 24 aligned contacts. Finally, choosing the light chain as the fingerprint, the algorithm picks out the heavy chain sequences as being similar, even though their sequence homology is low.

(b) *Globin/phycocyanin similarity*

Emboldened by the above results, we then undertook a far more rigorous test of the methodology and screened the sequence of the c-phycocyanin alpha chain (PHA 2) against the fingerprint library. Based on the relative scoring over a range of gap and insertion parameters, we found that the most compatible structure is sea hare myoglobin (1mba), and the resulting alignment of the sequences after the first pass is shown below with 1mba on top:

```
                SLSAAEADL AGKSWAPV FA NKNA N  GLD FLVALFEKF PDSANFFADFKGKSV ADIKASP KL RDVSSRIFTRLN 1mba
                .||..|... . . .... . . . .  .. ..  ... . .. ........ |.. .| .. . |. || .   . |..
MKTPLTEAVATADSQGRFLSSTELQVAFGR FRQASASLDAAKALSSKANSLAQGAVNAVYQK FPYTTQMQGKNFASD QRGKDKCARD IGYYI RIV PHA2

E FVNNAANAGKMSAM LSQFAK EHVGFGVGSA   QFENVRSM F PGFVASVAAPPAGADAAWTKLFGLIIDALKAAGA                    1mba
. .. .  . ..... . .|. .. . . .  ..| .. . . . .  | .. .. |.....  . .|`.||.
TYCL VAGGTGPLDDYLIGGLAEINR  T FDLSPSWYVEALKYIKANHGL  SGDPAVE ANSYID  Y AI NALS                    PHA2
```

The phycocyanin sequence for this alignment in the frozen approximation has an energy of −26·2 kT. After the second pass, the energy decreases to −32·9 kT and, as shown below with 1mba on top, the alignment shifts substantially.

```
                          SLSAAEADLAGKSWAPVFANK NANGLDFLV ALFEKFPDSANFFADFKGKSV ADIKASPKLRDVSSRIF 1mba
                          ...|.|. ..  . . ...| |.. ..  |  ...|    ........ |.. .| .. .  ..... .
MKTPLTEAVATADSQGRFLSSTELQVAFGRFRQASAS LD  AAKALSSKANSL AQGAVNAVYQK    FPYTTQMQGKNFASD QRGK  DKCARD I PHA2

TRLNEFVN  NA ANAGKMSA MLSQFAKEHVGFGVGSAQFENVRSMFPGFVASVA AP PAGADAAWT  KLFGLIIDALKAAGA                 1mba
  ....|.  ..   . .... ..  .|  .......   ..  .|. .. .. .. .. . . |.|..  :.....|.||.
GYYIRIVTYCLVAGGTGPLDDYLIGGLA  EINRTFD   LS  PSWYVEALKYIKANHGL SGDPAVEANSYIDYAINALS                 PHA2
```

Since the properties of sperm whale myoglobin (1mbd) have been more extensively studied, we then ran 1mbd against the entire sequence database and, as indicated in Figure 3(b), the phycocyanin sequence is confirmed as being compatible with the globin fold. Thus, the method correctly identifies the correspondence of the phycocyanin–globin fold, even though the two classes of molecules have little sequence homology (Pastore & Lesk, 1990). The resulting structural alignment is quite close to that obtained by Pastore & Lesk (1990), based on knowledge of the crystal structures of both molecules. Once again, for more distantly related sequences, the importance of the second pass where the interaction partners are those in the actual sequence is demonstrated.

Furthermore, as Figure 3(b) indicates, most of the top 500 scores are globin structures with even the most distantly related families represented. There are some sequences of unknown structure intermixed with the globin/phycocyanins. Some of these assignments may be wrong, others may be correct, and others may identify elements of supersecondary structure. Our strategy at this point is to proceed as in the phycocyanin case by screening putative sequence alignments against the entire structure library and then subjecting the best fitting structure to subsequent lattice stability analysis.

(c) *Identification of TIM folds*

Using the identical methodology, as shown in Figure 3(c), six of the ten families of the TIM fold (Farber & Petsko, 1990) appear in the top 50 scores against the chicken triose phosphate isomerase fingerprint. These include RUBISCO, the alpha subunit of tryptophan synthase, fructose biphosphate aldolase and enolase. Other sequences whose relationship to TIM is unclear were also found. Interestingly, depending on the buried/exposed pattern, the method seems capable of distinguishing between the monomeric and dimeric forms of the TIM type molecules.

Finally, we note that in preliminary applications, the method seems capable of identifying the structural homology of the cloverleaf fold of interleukin 1 and soybean trypsin inhibitor; this is another example of cases of identical structure, but little, if any, sequence homology. A detailed analysis will be presented in a future publication.

4. Conclusion

In this paper, we have presented a set of tools that allows the building of a low resolution model of the three-dimensional structure, provided that there is even a very remotely similar fold in the library of known protein structures. The fingerprint matching method provides a fast and very sensitive means for

detecting possible structural similarities. By being able to identify similar structures with no obvious relationship between sequences, it goes far in the direction of being a topology predictor. Subsequent model building using a lattice model provides a fast and automated way of further checking the alignment and is able to converge upon a low resolution prediction of tertiary structure.

There are a number of apparent extensions and applications of the methodology introduced here. For example, on a very crude level, if a sequence does not match a particular structure in the fingerprint library, we still find that the algorithm can classify proteins as α/β or mixed α/β. If it proves possible to employ the algorithm as a supersecondary structure predictor, one can imagine a scheme where one first classifies the protein by type, then identifies elements of supersecondary structure and finally uses the lattice approach to construct a prediction of the full tertiary structure. While the method undoubtedly requires refinement, we are very optimistic about the promise and utility of the combined fingerprint/lattice methodology, not only for the inverse folding problem, where it is working quite well, but also for the more general and more complicated problem of tertiary structure prediction.

We thank Drs W. Beers, J. Dyson, A. Finkelstein, R. Lerner, C. Sander, P. Wright and L. Walters for stimulating discussions, and M. Pique and Y. J. Chen for help with preparation of the Figures. This work was supported in part by grant no. GM-37408 of the Division of General Medical Sciences of the National Institutes of Health and the Joseph Drown Foundation.

References

Altschul, S. F. & Lipman, D. J. (1990). Protein database searches for multiple alignments. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 5509–5513.

Adman, E. T., Turley, S., Bramson, R., Petratos, K., Banners, D., Tsernoglous, D., Beppu, T. & Watanabe, H. (1989). A 2·0 Å structure of the blue copper protein (Cupredoxin) from *Alcaligenes faecalis* S-6. *J. Biol. Chem.* **264**, 87–99.

Argos, P. (1987). A sensitive procedure to compare amino acid sequences. *J. Mol. Biol.* **193**, 385–396.

Baker, E. N. (1988). Structure of azurin from alcaligenes denitrificans refinement at 1·8 Å resolution and comparison of the two crystallographically independent molecules. *J. Mol. Biol.* **203**, 1071–1095.

Bernstein, F. C., Koetzie, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bairoch, A. (1989). Ph.D. Thesis, University of Geneva.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Chothia, C. (1982). Evolution of proteins formed by beta-sheets. *J. Mol. Biol.* **160**, 309–323.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature (London)*, **319**, 199–302.

Farber, G. K. & Petsko, G. A. (1990). The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* **15**, 228–234.

Finkelstein, A. V. & Reva, B. A. (1991). A search for the most stable folds of protein chains. *Nature (London)*, **351**, 497–499.

Furey, W., Jr, Wang, B. C., Yoo, C. S. & Sax, M. (1983). Structure of a novel Bence–Jones protein (1rhe) fragment at 1·6 Å resolution. *J. Mol. Biol.* **167**, 661–692.

Getzoff, E. D., Tainer, J. A., Stempien, M. M., Bell, G. I. & Hallewell, R. A. (1989). Evolution of CuZn superoxide dismutase and the Greek key beta-barrel structural motif. *Proteins*, **5**, 322–336.

Godzik, A. & Sander, C. (1989). Conservation of residue interactions in a family of Ca-binding proteins. *Protein Eng.* **2**, 589–596.

Godzik, A. & Skolnick, J. (1992). Modularity and specificity of globular protein tertiary contact maps. *Proc. Nat. Acad. Sci., U.S.A.* In the press.

Gribskov, M., MacLachlan, A. D. & Eisenberg, D. (1987). Profile analysis detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4359.

Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1·6 Å resolution. *J. Mol. Biol.* **169**, 521–563.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Holm, L., Saraste, M. & Wikström, M. (1987). Structural models of the redox centres in cytochrome oxidase. *EMBO J.* **6**, 2819–2823.

Kawula, T. H., Spinola, S. M., Klapper, D. G. & Cannon, J. G. (1987). Localization of a conserved epitope and an azurin-like domain in the h.8 protein of pathogenic neisseria. *Mol. Macrobiol.* **1**, 179–185.

Lesk, A. M. (1991). *Protein Architecture. A Practical Approach.* IRL Press, New York.

Musil, D., Zucic, D., Turk, D., Engh, R. A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katunuma, N. & Bode, W. (1991). The refined 2·15 Å X-ray crystal structure of human liver cathepsin b: the structural basis for its specificity. *EMBO J.* **10**, 2321–2330.

Needleman, S. B. & Wünsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Nikishawa, K., Ooi, T., Ysogai, Y. & Saito, N. (1972). Tertiary structure of proteins. I. Representation and computation of conformations. *J. Phys. Soc. Jpn.* **32**, 1311–1337.

Pastore, A. & Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins*, **8**, 133–155.

Phillips, D. C. (1970). The development of crystallographic enzymology. *Biochem. Soc. Symp.* **30**, 11–28.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. *J. Mol. Biol.* **193**, 775–791.

Protein Sequence Database Release 19, October 1991, EMBL Data Library, Heidelberg, Germany.

Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data. *Proteins*, **3**, 71–84.

Scharf, M. (1989). Diplomarbait, University of Heidelberg.

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.

Skolnick, J. & Kolinski, A. (1989). Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.* **40**, 207–235.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, **250**, 1121–1125.

Skolnick, J. & Kolinski, A. (1991). Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* **221**, 499–531.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1991). Prediction of progress at last. *Nature (London)*, **354**, 105–105.

Van Beeumen, ?., Van Bun, S., Canters, G. W., Lommen, A. & Chothia, C. (1991). The structure homology of amicyanin from *Thiobacillus versutus* to plant plastocyanins. *J. Biol. Chem.* **266**, 4869–4877.

Waterman, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.* **46**, 473–500.

Edited by F. E. Cohen