Journal of Computer-Aided Molecular Design, 7 (1993) 439–456
ESCOM

J-CAMD 208

Protein fold recognition

David Jones and Janet Thornton

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, U.K.

Received 24 November 1992
Accepted 26 February 1993

Key words: Proteins; Protein structure; Protein folding; Tertiary structure prediction; Computer algorithms

SUMMARY

An important, yet seemingly unattainable, goal in structural molecular biology is to be able to predict the native three-dimensional structure of a protein entirely from its amino acid sequence. Prediction methods based on rigorous energy calculations have not yet been successful, and best results have been obtained from homology modelling and statistical secondary structure prediction. Homology modelling is limited to cases where significant sequence similarity is shared between a protein of known structure and the unknown. Secondary structure prediction methods are not only unreliable, but also do not offer any obvious route to the full tertiary structure. Recently, methods have been developed whereby entire protein folds are recognized from sequence, even where little or no sequence similarity is shared between the proteins under consideration. In this paper we review the current methods, including our own, and in particular offer a historical background to their development. In addition, we also discuss the future of these methods and outline the developments under investigation in our laboratory.

INTRODUCTION

A fairly recent and exciting development in protein structure prediction is the approach of recognizing whole protein folds from sequence. Facilitated by the current wide availability of powerful computers, a number of groups, including our own, have developed quite successful techniques for aligning sequences with structures, and using these alignments to detect the native fold of a protein sequence from a set of alternatives. In this review we shall be taking a historical view of this work, putting in context the earlier work on protein structure prediction, and showing how this work has influenced the more recent developments.

Current success in predicting a protein's native fold from first principles is extremely limited, and consequently other 'heuristic' methods have been found. An obvious approach to solving vastly complex systems of equations (which the protein folding equations must surely be) is to merely observe the macroscopic properties exhibited by a wide range of different final solutions. In this case, a sensible approach is to analyse the final folded states of different proteins statistically. Amongst the attempts at analysing the relationship between protein sequence and structure statistically were those of Chou and Fasman [1] and Garnier et al. [2]. These attempts were strictly aimed at predicting the secondary structure of proteins. The basic idea behind these techniques is to assign a structural 'propensity' to either individual residues (e.g. Chou and Fasman) or short sequence segments (Garnier et al.). For a brief review of these purely statistical approaches see Taylor [3], or for a more extensive coverage see Fasman [4].

SUPER-SECONDARY STRUCTURE PATTERNS

All the above pattern matching methods are concerned with detecting simple patterns primarily designed to predict secondary structure. The trend in protein sequence pattern matching has moved towards the construction of more complex templates, capable of matching higher levels of structure than α-helical or β-sheet regions. Given that long-range interactions play an important part in the directing of protein folding, even at the purely secondary structural level, the necessity for these complex patterns is reasonably obvious. Nagano [5] first described the use of a super-secondary structural motif, where the βαβ unit was analysed. Following on from this work, Nagano [6] extended the algorithm to a generalized structure prediction system. This approach splits the sequence into pentapeptides to reduce the number of degrees of freedom in the folding simulation. The folding is further constrained by considering the packing in a 2D matrix (3 × 11 boxes) rather than the 3D atomic coordinate space. The nub of the method is simple in that each pentapeptide is labelled as being α or β depending on the standard secondary structure prediction probability for each. Likely βαβ units are then located by considering pentapeptide patterns that neighbour strongly predicted α or β segments, with an appropriate distance filter that excludes β–α/α–β pairs that are too far apart on the grid. An important part of this method was a combinatorial analysis of all the possible permutations of predicted structural segments, which is analogous to scanning through all the possible 3D packings given a mixture of well-defined and ill-defined structural units, except that in this case packing is performed on a 2D grid.

Taylor and Thornton [7] attempted to improve the accuracy of secondary structure prediction by constructing templates capable of detecting super-secondary structural elements, or more specifically (but not exclusively) the βαβ unit. Using 62 examples of the βαβ unit from the Brookhaven database, an ideal secondary structure sequence template was constructed [8]. This ideal βαβ pattern was matched at each residue position of the test sequence matching the template profiles to the Garnier secondary structure prediction probability profiles [2] and a score was calculated. Different length variants of the ideal template were created by scaling the master template so as to accommodate the length variations observed in the available examples. Apart from the statistical sequence template, templates were also constructed for matching patterns of hydrophobicity – one template scored highly for buried β regions, the other for α-helical regions. The strongest fitting template was selected, and other matching templates were selected according to various rules (for example forbidding overlapping α and β regions). On a test set of 16 β/α proteins, a prediction score of 70% was achieved, bettering the raw GOR secondary structure prediction technique by some 7.5%.

The success of these simple super-secondary structure recognition methods was unfortunately somewhat limited. Whilst these methods were very capable of identifying the target motifs, they had an unfortunately high false–positive rate. This was an inevitable result of their reliance on secondary structure pattern recognition. Clearly at the level of secondary structure and hydrophobicity, a βαβ super-secondary structural motif is difficult to distinguish from other non-specific βαβ patterns. A sensible way to progress from these super-secondary structure methods is towards the recognition of larger more specific folding patterns, in particular a complete chain fold.

SEQUENCE TEMPLATES AND PROFILES

Initially attempts at recognizing complete folds focused on recognizing the physicochemical properties that were conserved across a family of aligned sequences. In a strict sense, these methods do not actually recognize folds, but detect remote sequence relationships.

As an extension to the work on βαβ templates, Taylor [9] went on to produce a generalized consensus template method. The first major improvement to the original template method of Taylor and Thornton was to move over to 2D templates that could match more than one physicochemical criterion at each alignment position. The second major improvement was to contrive a means for generating the templates automatically, given a suitably well-defined sequence alignment. The method starts with a seed alignment, generally based on available structural information. A consensus pattern is created from this initial alignment such that each alignment position contains a count of each of the 20 amino acids. Rather than recording the number of each of the 20 amino acids observed at each position, the observed amino acid identities are used to pick a minimal covering class of amino acid (i.e. the smallest subset that contains all the residues observed at a particular alignment position) from a Venn diagram [10]. This Venn diagram comprises three major sets: Hydrophobic, Polar and Small (with subsets Aromatic, Aliphatic, Tiny, Charged and Positive). This final template of amino acid property sets could then be aligned against new sequences in order to identify a possible relationship.

Pearl and Taylor [11] used the above consensus template program suite to identify common features in the retroviral protease and aspartyl protease families. The method was able to detect the few conserved residues that formed the proposed active site even though the sequences showed almost no similarity when compared with standard pairwise alignment methods.

Other workers have developed similar template-based methods [12–14]. In particular, the profile method of Gribskow et al. [13] has proven to be very popular due to its reliable statistical properties.

Whilst these methods have proven to be very successful at detecting relationships between proteins with similar biological function, where functional constraints limit the degree of sequence divergence, in cases where no functional constraints exist, these methods have been unsatisfactory. A good example of this is the case of the globins and the phycocyanins. The folding patterns of these two functionally dissimilar families are remarkably similar [15], yet none of the sequence-based template methods have been able to detect this match [12]. The conclusion we must make on the basis of this observation is that, as might be expected, sequence template methods are only capable of exploiting weak sequence similarities and are not capable of looking beyond the limits imposed by divergent sequence evolution. Despite the limitations of these methods, they have certainly helped spawn the more modern approaches of fold recognition, and remain a very useful and reliable tool in protein sequence analysis.

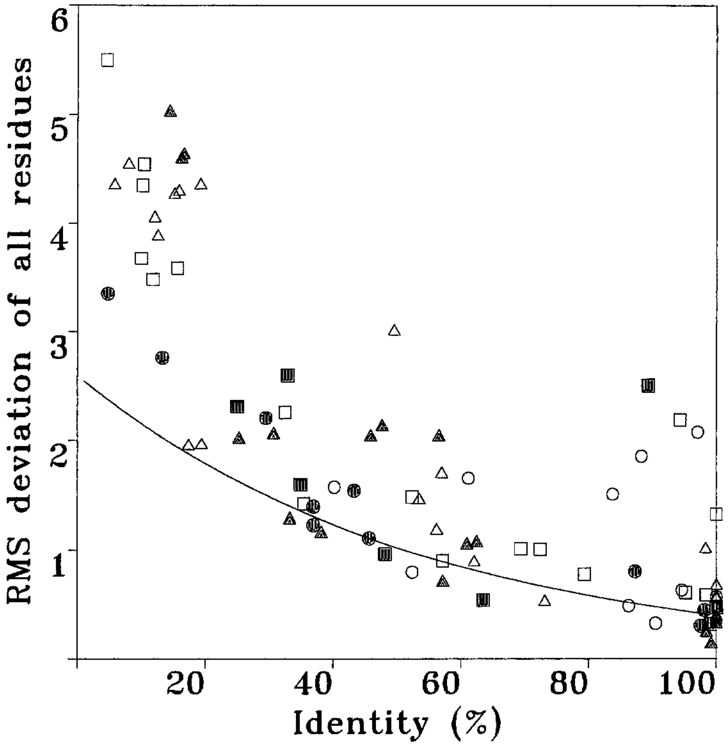

Fig. 1. The Root Mean Square Deviation (RMSD) between 90 structurally aligned pairs of protein chains plotted against the residue identity scores [19]. Proteins are classified into mostly-α (○), mostly-β (△), and αβ classes (□). Pairs for which the symbol is filled have been refined and are resolved to at least 2.0 Å.

FOLD RECOGNITION

Many fragments of evidence point towards there being a limited number of *naturally occurring* protein folds. If we consider a chain 50 residues long and assume that each residue has seven accessible conformational states [16], we might naively calculate the number of possible main chain conformations as 7^{50} ($\approx 10^{42}$). Clearly most of these conformations will not be stable folds, many will be trivially similar, and many will not be even physically possible. In order to form a compact globular structure a protein chain necessarily has to form regular secondary structures [17], and it is this constraint, along with the constraints imposed from a requirement to effectively pack the secondary structures formed, that limits the number of stable conformational states for a protein chain. In addition to the constraints imposed by physical effects on protein stability, there are evolutionary constraints on the number of folds. Where do new proteins come from? The answer according to Doolittle [18] is from other proteins. In other words, the folding patterns we observe today are the result of the evolution of a set of ancestral protein folds. Despite the fact that much is now known about the effects of a small number of sequence mutations on the final tertiary structure, we have only circumstantial evidence about the effects of extensive sequence divergence. Figure 1 shows the relationship between sequence similarity and structural similarity for 90 pairs of protein chains, after structural alignment [19]. It is quite clear that the effects of

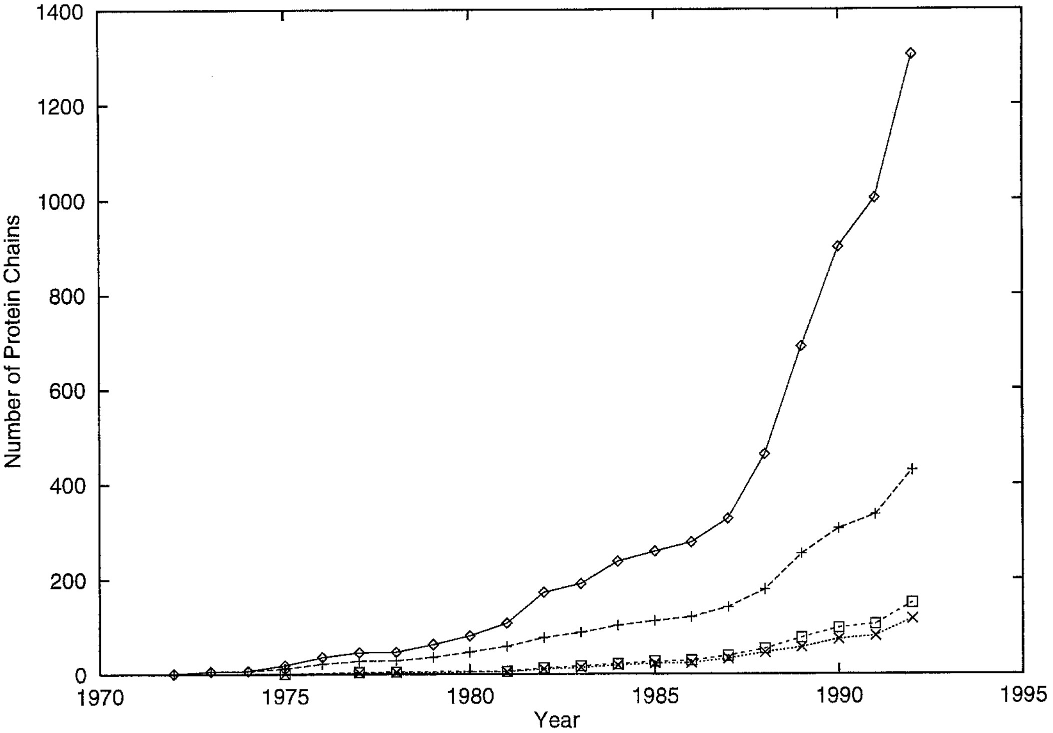

Fig. 2. The increase [22] in the numbers of protein structures (determined by X-ray crystallography and NMR spectroscopy) deposited in the Brookhaven database [8]. The top line (◇) gives the total number of structures. The line below (+) shows the increase in numbers of nonidentical structures (< 98% sequence identity) and the lower two lines are the numbers of nonhomologous (□) (< 35% sequence identity) and nonanalogous (X) (< 35% identity and no evident structural similarity) folds. At present only 10% of the deposited structures have new folds.

sequence divergence become quite complex as the degree of sequence similarity approaches the region of 20–30% sequence identity. At this point it no longer remains clear whether the proteins in question share an evolutionary relationship, or whether they have evolved from unrelated ancestors and have evolved *convergently* towards a particular stable chain fold. This latter case now appears to be the most likely explanation for the similarities and distinct differences between the families of TIM barrel enzymes [20,21]. As a further illustration of the limited number of folds, Fig. 2 shows the rate at which new structures are being deposited in the Brookhaven database [22], and the number of these structures found to have similar sequences and folds.

A limited number of folds and the recurrence of folds in proteins which share no significant sequence similarity offer a 'short cut' to protein tertiary structure prediction. It is quite obvious that it is impractical to attempt tertiary structure prediction by searching a protein's entire conformational space for the minimum energy structure. That is not to say that more 'intelligent' search procedures will not be found which are capable of quickly eliminating impossible regions of conformational space (future developments of stochastic search algorithms such as simulated annealing). If we know that there are as few as 1000 protein families [23] and consequently no more than 1000 possible protein folds, then the intelligent way to search a protein's conforma-

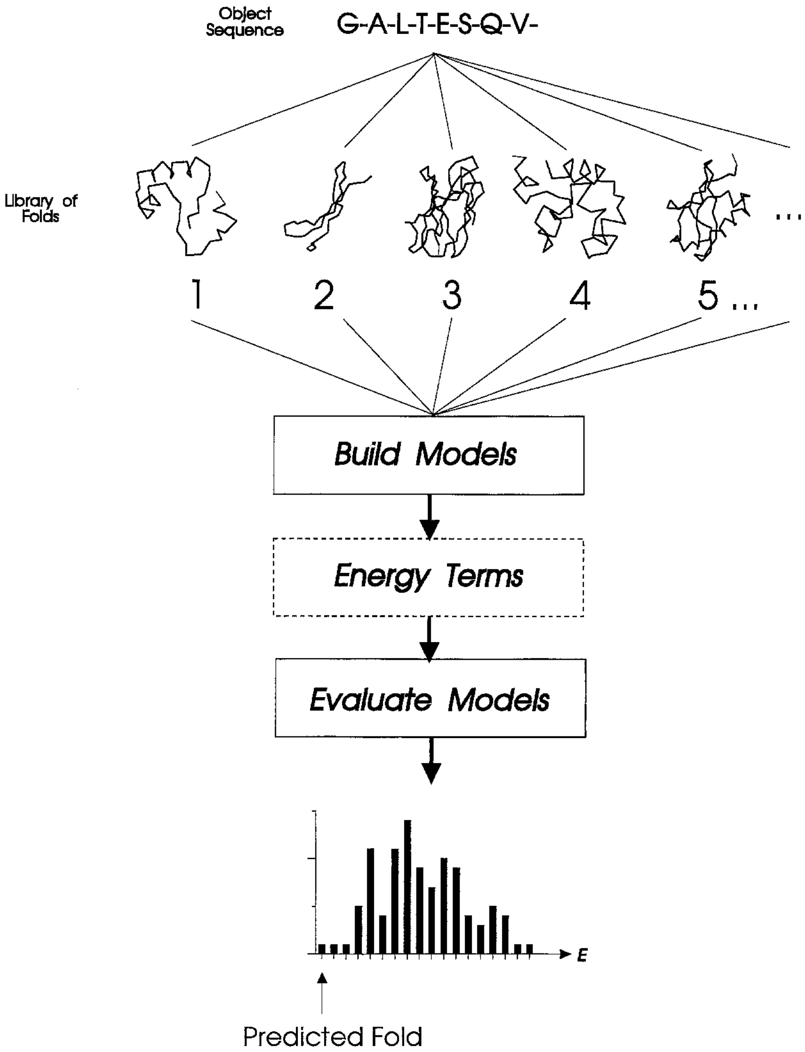

Fig. 3. A conceptual outline of fold recognition as a solution to the protein folding problem. A given sequence is fitted to the backbones of known structures, and the goodness-of-fit in each case is evaluated by one of the many available model evaluation procedures.

tional space would be to simply consider only those regions that correspond to this predefined set. This is analogous to the difference between a 'free answer' exam paper, requiring the writing of an essay on a given topic, and an exam merely requiring multiple-choice questions to be answered. Clearly a person with no knowledge of the subject at hand has a much greater chance of achieving success with the multiple-choice paper than with the free answer paper. By considering only a few alternative conformations, we are also less reliant on accurate models of the energetics of protein stability. Even if we were able to search a protein's entire conformational space, and even if the native fold did correspond to the global energy minimum, then we would still be stuck because none of the existing empirical force fields are good enough models of a protein's free energy. Furthermore, the hydrophobic effect, which is possibly the most important consideration in determining a protein's tertiary fold, is an entropic effect requiring the accurate simulation of bulk solvent surrounding the protein chain. Fortunately, it seems that even simple models of the hydrophobic effect alone are good enough to identify the native fold of a protein from a small number of plausible alternatives.

Recent attempts at protein fold recognition have considered the following problem. Given a sequence, and a library of folds, can the correct native fold be identified? Figure 3 illustrates our own interpretation of the fold recognition problem. Some groups have rephrased the question along the lines suggested by Drexler [24] and have tackled the *inverse folding problem.* In the case of the inverse folding problem, the problem is to extract from a set of sequences those sequences that are compatible with a single given *structure.* Despite some practical differences between these two formulations of the same problem, both depend on methods for evaluating the compatibility of a sequence with a fold.

EVALUATING STRUCTURAL MODELS

The inability of standard atomic force fields to detect misfolded proteins was first demonstrated by Novotny and Karplus [25]. Their test problem was very simple, and yet is a good illustration. In this study, the sequences of myohemerythrin and an immunoglobulin domain of identical length were exchanged. Both the two native structures, and the two 'misfolded' proteins were then subjected to energy minimization, using the CHARMm [26] force field. The results were somewhat surprising in that it was impossible to distinguish between the native and misfolded structures on the basis of the calculated energy sums. Novotny and Karplus correctly surmised that the reason for this failure was the neglect of solvation effects in the force field. In a later study [27] the force field was modified to approximate the effects of solvent and in this case the misfolded structures could be identified. The work of Novotny and Karplus encouraged several studies into effective methods for evaluating the correctness of protein models, which will now be briefly reviewed.

Eisenberg and McLachlan [28] were able to distinguish correct models from misfolded models by using a simple solvation energy model alone. By calculating a solvation free energy for each amino acid type and by calculating the degree of solvent accessibility for each residue in a given model structure, the correctly folded models were clearly distinguished from the misfolded.

Baumann et al. [29] also used a solvation term to recognize misfolded protein chains, along with a large number of other general statistical properties of sequences forming stable protein folds. Holm and Sander [30] have recently proposed another solvation model, which appears to be very able at detecting misfolded proteins, even those proteins that have shifts of their sequence on their correct native structure. Interestingly enough, a sequence–structure mismatch can quite easily occur not just in theoretically derived models, but even in crystallographically derived models. For example, one of the xylose isomerase structures in the current Brookhaven database has in part a clearly mistraced chain. Such errors can be detected by use of a suitable solvation-based model evaluation procedure.

Several groups have used statistically derived pairwise potentials to identify incorrectly folded proteins. Using a simplified side chain definition, Gregoret and Cohen [31] derived a contact preference matrix and attempted to identify correct myoglobin models from a set of automatically generated models with incorrect topology, yet quite reasonable core packing.

Hendlich et al. [32] used a set of potentials of mean force, first described by Sippl [33] to not only correctly reject the misfolded protein models of Novotny and Karplus, but also to identify the native fold of a protein amongst a large number of decoy conformations generated from a database of structures. The protein sequence of interest was blindly fitted to all contiguous structural fragments taken from a library of highly resolved structures, and the contact energy terms were summed in each case. For example, consider a protein sequence of 50 residues being fitted to a structure 100 residues long. The structure would offer 51 possible conformations for this sequence, starting with the sequence being fitted to the first 50 residues of the structure, and finishing with the sequence being fitted to the last 50. Taking care to eliminate the test protein from the calculation of potentials, Hendlich et al. correctly identified 41 out of 65 chain folds. Using factor analysis, Casari and Sippl [34] have found that the principal component of their potentials of mean force is a hydrophobic potential of simple form. This principal component potential alone is found to be almost as successful as the full set of potentials in identifying correct folds.

In a very similar study to that performed by Hendlich et al., Crippen [35] used a simple discrete contact potential to identify a protein's native fold from all contiguous structural fragments of equal length extracted from a library of structures. The success rate (45 out of 56) in this case was marginally higher than that of Hendlich et al., because the contact parameters were optimized against a 'training set' of correct and incorrect model structures. Maiorov and Crippen [36] improved these results by using a continuous contact potential, with the new contact function correctly identifying virtually all chain folds defined as being 'compact'.

Both the work of Hendlich et al. and Crippen demonstrates a very restricted example of fold recognition, whereby sequences are matched against suitable sized contiguous fragments in a template structure. A much harder recognition problem arises when more complex ways of fitting a sequence to a structure are considered, i.e., by allowing for relative insertions and deletions between the object sequence and the template structure. Suitable treatment of insertions and deletions is essential to a generalized method for protein fold recognition.

Ponder and Richards [37]

The first true example of a fold recognition attempt was the template approach of Ponder and Richards [37]. They concerned themselves with the inverse folding problem in that they tried to enumerate sequences that would be compatible with a given backbone structure. The evaluation potential was a simple van der Waals potential, and so models were effectively scored on the degree of overlap between side chain atoms. A further requirement was for the core to be well packed, which was achieved by considering the conservation of side chain volume. In order to fit the side chains of a given sequence onto the backbone, an exhaustive search was made through a 'rotamer library' of side chain conformations. If after searching rotamer space the side chains could not be fitted successfully into the protein core, then the sequence was deemed incompatible with the given fold. As a sensitive fold recognition method, however, this method was not successful. Without allowing for backbone shifts, the packing requirement of a given protein backbone was found to be far too specific. Only sequences very similar to the native sequence could be fitted successfully to the fixed backbone.

Bowie et al. [38]

A rather more successful attempt at fold recognition was made by Bowie et al. [38]. The first stage of this method involves the prediction of residue accessibility from multiple sequence alignments, which is itself another interesting recent development (see later). In essence, alignment positions with high average hydrophobicity and high conservation are predicted to be buried and relatively polar variable positions are predicted to be exposed to solvent. The degree of predicted exposure at each position of the aligned sequence family is then encoded as a string. This string is then matched against a library of similarly encoded strings, based, however, not on predicted accessibilities but on *real* accessibilities calculated from structural data. Several successful recognition examples were demonstrated using this method. Of particular note was the matching of an aligned set of Ef Tu sequences with the structure of flavodoxin. The similarity between Ef Tu and flavodoxin is not readily apparent even from structure [22] and so this result is quite impressive.

Bowie et al. [39]

Bowie et al. [39] have attempted to match sequences to folds by describing the fold not just in terms of solvent accessibility, but in terms of the *environment* of each residue location in the structure. In this work, the environment is described in terms of local secondary structure (three states: α, β and coil), solvent accessibility (three states: buried, partially buried and exposed), and the degree of burial by polar rather than apolar atoms. The environment of a particular residue thus defined tends to be more highly conserved than the identity of the residue itself, and so the method is able to detect more distant sequence-structure relationships than purely sequence-based methods. The method has also been applied to the evaluation of protein models [40]. The authors describe this method as a 1D–3D profile method, in that a 3D structure is encoded as a 1D string, which can then be aligned using traditional dynamic programming algorithms. Bowie et al. have applied the 1D–3D profile method to the inverse folding problem and have shown that the method can indeed detect remote matches but, in the cases shown, the hits have still retained some sequence similarity with the search protein, even though in the case of actin and the 70 kD heat-shock protein the sequence similarity is very weak [41]. Environment-based methods appear to be incapable of detecting structural similarities between extremely divergent proteins and between proteins sharing a common fold through convergent evolution – environment only appears to be conserved up to a point. Consider a buried polar residue in one structure that is found to be located in a polar environment. Buried polar residues tend to be functionally important residues, and so it is not surprising then that a protein with a similar structure but with an entirely different function would choose to place a hydrophobic residue at this position in an apolar environment. A further problem with environment-based methods is that they are sensitive to the multimeric state of a protein. Residues buried in a subunit interface of a multimeric protein will not be buried at an equivalent position in a monomeric protein of similar fold. In a rather roundabout way, the same authors went on to use this method to successfully evaluate protein models [40] and with a commendable degree of frankness demonstrated that the method was capable of detecting a previously identified chain tracing error in a structure solved in their own laboratory.

Finkelstein and Reva [42]

Finkelstein and Reva [42] have used a simplified lattice representation of protein structure for their work on fold recognition. The problem they consider is that of matching a sequence to one of the 60 possible eight-stranded β-sandwich topologies. Each strand has three associated variables: length, position in the sequence and spatial position in the lattice Z direction. The force field used by Finkelstein and Reva includes both short-range and long-range components. The short- range component is simply based on the β-coil transition constants for single amino acids, similar in many respects to the standard Chou–Fasman [1] propensities. The long-range interaction component has a very simple functional form. For a pair of interacting (contacting) residues, it is defined simply as the sum of their solvent transfer energies, as calculated by Fauchere and Pliska [43].

The configurational energy of the eight strands in this simple force field is minimized by a simple iterative method. At the heart of the method is a probability matrix (a 3D matrix in this case) for each of the strands, where each matrix cell represents one triplet of the strand variables, i.e. length, sequence position and spatial position. The values in each cell represent the probability of observing the strand with the values associated with the cell. The novel aspect of this optimization strategy is that the strands themselves do not physically move in the force field, only the probabilities change. At the start of the first iteration the strand coordinate probabilities are assigned some arbitrary value, either all equal, or set close to their expected values (for example, the first strand is unlikely to be positioned near the end of the sequence). A new set of probabilities is then calculated using the current mean field and the inverse Boltzmann equation. As more iterations are executed it is to be hoped that most of the probabilities will collapse to zero, and that eventually a stable 'self-consistent' state will be reached. Finkelstein and Reva found that the most probable configurations corresponded to the correct alignment of the eight-stranded model with the given sequence, and that when the process was repeated for each of the 60 topologies, in some cases the most probable configuration of the native topology had the highest probability of all.

The simplicity of the lattice representation and the uncomplicated force field are critical to the success of this method. A more detailed interresidue potential would prevent the system from reaching a self-consistent state, and would be left either in a single local minimum or more likely oscillating between a number of local minima. In addition, whilst it is quite practical to represent β-sheets on a lattice, it is not clear how α-helices could be reasonably represented. It will be interesting to see whether this method can be extended to classes of protein structure other than the all-β class.

Jones et al. [44]

The method we have developed in collaboration with W.R. Taylor [44] has something in common both with the method of Bowie, Lüthy and Eisenberg [39] and that of Finkelstein and Reva [42]. Despite the obvious computational advantages of using residue environments, it is clear that the fold of a protein chain is governed by fairly specific protein-protein and protein-solvent atomic interactions. A given protein fold is therefore better modelled in terms of a 'network' of pairwise interatomic energy terms, with the structural role of any given residue described in terms of its interactions. Classifying such a set of interactions into one environmental class such as 'buried α-helical' will inevitably result in the loss of useful information, reducing the *specificity* of sequence-structure matches evaluated in this way. Put simply, one amphipathic helix is much like any other amphipathic helix. Ideally, we would like to match a sequence to a structure by considering the plethora of detailed pairwise interactions, rather than averaging them into a crude environmental class. However, incorporation of such nonlocal interactions into

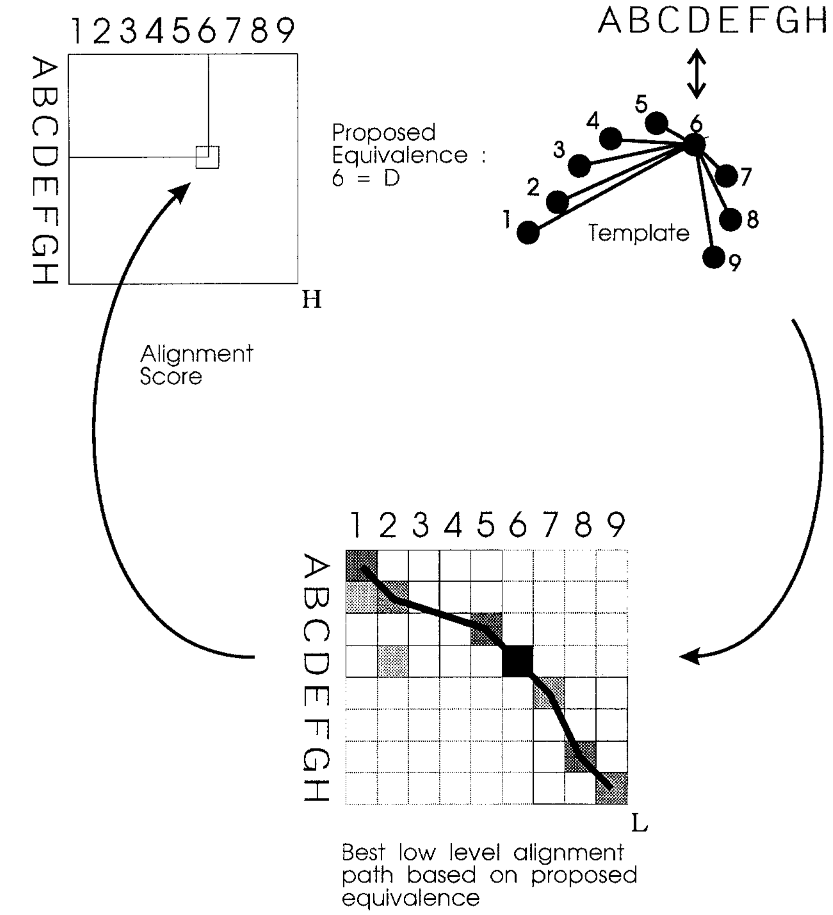

Fig. 4. The core of our threading algorithm, whereby the residues in the sequence are matched against the structural environments of residues in the template fold. The sequence elements are labelled A-H, and the template elements 1–9. The steps illustrated show how a single cell of the high-level matrix is calculated (see Ref. 46).

standard alignment methods such as the algorithm of Needleman and Wunsch [45], has hitherto proved computationally impractical.

We have applied a novel dynamic programming algorithm (now commonly known as 'double' dynamic programming) to the problem of aligning a given sequence with the 'real' coordinates of a structure , taking into account the detailed pairwise interactions, a process which we call *optimal sequence threading*. The requirement here to match pairwise interactions relates to the requirement of structural comparison methods. We define here the *potential environment* of a residue i as being the sum of all pairwise potential terms involving i and all other residues $j \neq i$. This is a similar definition to that of a residue's *structural environment*, as described by Taylor and Orengo [46]. In the simplest case, the structural environment of a residue i is defined as the set of all inter-Cα distances between residue i and all other residues $j \neq i$. Taylor and Orengo propose a novel dynamic programming algorithm for the comparison of residue structural environments, and we have used a derivation of this method for the effective comparison of residue potential environments. Figure 4 outlines the steps involved in the use of double dynamic programming for aligning sequences with structural templates.

Whilst our method is applicable to any form of pairwise potential, including all those which have been described in this review, we chose to use a set of statistically derived pairwise potentials similar to those described by Sippl [33]. Using the formulation of Sippl, we have constructed short- (sequence separation, $k \leq 10$), medium- ($11 \leq k \leq 30$) and long- ($k > 30$) range potentials between the following atom pairs: $C\beta \rightarrow C\beta$, $C\beta \rightarrow N$, $C\beta \rightarrow O$, $N \rightarrow C\beta$, $N \rightarrow O$, $O \rightarrow C\beta$ and $O \rightarrow N$. For a given pair of atoms, a given residue sequence separation, and a given interaction distance, these potentials provide a measure of energy, which relates to the probability of observing the proposed interaction in native protein structures. Our potentials differ from those proposed by Sippl in the following ways. Firstly, interactions beyond 10 Å are ignored. We have found that these interactions are not residue-specific and are determined simply by solvation effects. In addition, the longer-distance interactions are biased towards the larger proteins in the set used to calculate the potentials. Instead of the long-distance terms, we use a 'solvation potential'. This potential simply measures the frequency with which each amino acid species is found with a certain degree of solvation, approximated by the residue solvent-accessible surface area. We define the solvation potential for amino acid residue a as follows:

$$\Delta E^{a}_{solv.}(r) = -kT \ln[\frac{f^{a}(r)}{f(r)}] \qquad (1)$$

where r is the percentage residue accessibility (relative to residue accessibility in the GGXGG fully extended pentapeptide), $f^{a}(r)$ is the frequency of occurrence of residue a with accessibility r, and f(r) is the frequency of occurrence of all residues with accessibility r. Residue accessibilities were calculated using the program DSSP [47], applied to Brookhaven coordinate files [8]. For multimeric proteins, only the chains explicitly described in the coordinate files were taken into account.

A final point to note about our force field is that all pairwise terms involving loop residues were ignored, due to the fact that loop conformations tend not to be conserved even between closely related proteins, let alone distant relatives.

By dividing the empirical potentials into sequence separation ranges, specific structural significance may be tentatively conferred on each range. For instance, the short-range terms predominate in the matching of secondary structural elements. By threading a sequence segment onto the template of an α-helical conformation and evaluating the short-range potential terms, the possibility of the sequence folding into an α-helix may be evaluated. In a similar way, medium-range terms mediate the matching of super-secondary structural motifs, and the long-range terms, the tertiary packing.

We have found this method capable of detecting quite remote sequence-structure matches. In particular it has proven able to detect the previously mentioned similarity between the globins and the phycocyanins, and is therefore the first sequence analysis method to achieve this. Also of particular interest are the results for some $(\alpha\beta)_8$ (TIM) barrel enzymes and also the β-trefoil folds: trypsin inhibitor DE-3 and interleukin 1β, for example. The degree of sequence homology between different $(\alpha\beta)_8$ barrel enzyme families and between trypsin inhibitor DE-3 and interleukin 1β is extremely low (5–10%). As a consequence, sequence template methods have not proven able to detect these folds. It is therefore clear that new information beyond sequence similarity is being utilized in our method. Although the actin fold was found to be the best match when a hexokinase sequence was matched against the fold library, the separation between actin and the next fold in the list of folds sorted by 'threading energy' was close to zero. We might therefore surmise that the degree of structural similarity between hexokinase and actin represents the current limit to the sensitivity of the method. However, the sensitivity of the method does depend on the secondary structure content of the protein's native fold. For example, the method yields better results in detecting similarities between proteins in the all-α class of protein structure than in the αβ class, with the all-β class being the most difficult of all. Also proteins with very large relative insertions and deletions prove difficult to match. Of course, these problems are common to all protein-fold recognition methods.

WORK IN PROGRESS

Despite the success of the optimal-sequence threading method for fold recognition, there remain several exciting avenues for future study, two of which shall be described briefly. The first avenue is concerned with improving the sensitivity and selectivity of the method. The sensitivity needs to be improved in order that we can recognize even more remote relationships than those previously described. Results with the TIM barrels and the globins and phycocyanins are extremely encouraging, yet these folds are the folds with the most obvious and extensively shared structural similarities. Proteins sharing only limited structural similarity at the topological level, rather than the structural level are difficult to match. Other problems involve the matching of very small protein domains in which there is no large hydrophobic core, and consequently no easily identifiable solvation patterns. A good example of a structural similarity which seems to elude current fold recognition methods is the similarity observed between immunoglobulin-binding protein G and ubiquitin [48], see Fig. 5. Vriend et al. [49] have also noted a structural similarity between ferredoxin and ubiquitin, although in this case the hydrogen-bonding pattern is not so well conserved. Given that these proteins comprise few secondary structural elements, the number of possible topologies is small and it is therefore possible that this is a chance occurrence rather than the result of these proteins sharing a common ancestor.

Of course by far the most significant deficiency in structure prediction by fold recognition is simply that you can only predict that which you have previously observed.

MULTIPLE SEQUENCE INFORMATION

In some cases, rather than a single sequence, a whole family of related sequences is available for analysis. By multiple aligning of the sequence family, additional information may be obtained from the observed sequence-conservation patterns, and the location of insertions and deletions. A prime example of the power of multiple sequence data was the successful secondary-structure prediction of the cAMP-dependent kinases [50–52], although a recent prediction of the SH3 domain structure was not so successful [53]. At the most basic level, the likely location of loop regions in the sequence data can be derived by observing where insertions and deletions occur. This information alone will be of significant benefit to our fold recognition procedure. Beyond this, it is clear that fitting multiple sequences to a structure will increase both the sensitivity and selectivity of our method as the interaction energy between two positions in the structure would now depend on more than two residues at a time. Rather than matching one residue to

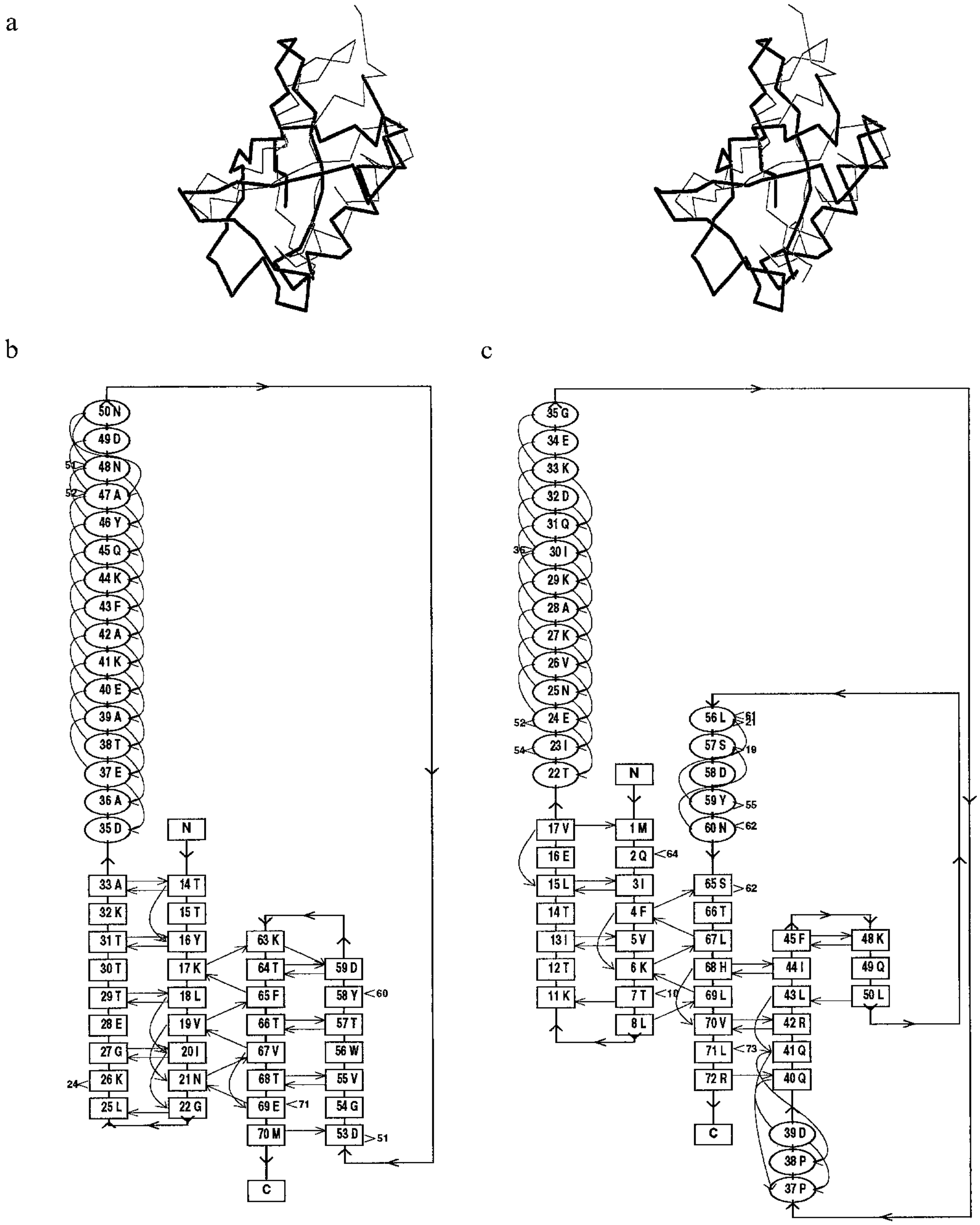

Fig. 5. Detecting the structural similarity between protein G and ubiquitin appears to be beyond the capabilities of current protein fold-recognition schemes. A stereodiagram of protein G (Brookhaven code IPGX, thin line), superposed onto ubiquitin (code 1UBQ, thick line) is shown in Fig. 5a. The RMSD is 3.36 Å calculated over 55 equivalenced Cα atoms. The similarity is very clearly seen in the hydrogen bonding diagrams [57], (Figs. 5b, c). In the hydrogen bonding diagrams, residues in sheets are drawn inside boxes and residues in helices are drawn inside ellipses. Another remote similarity has been reported between ubiquitin (thick line) and 2Fe-2S ferredoxin (code 1FXI, thin line) shown in Fig. 5d. In this case the RMSD for Cα atoms is 3.28 Å over 63 equivalences. The similarity is again apparent in the hydrogen bonding diagrams (Figs. 5b, e).

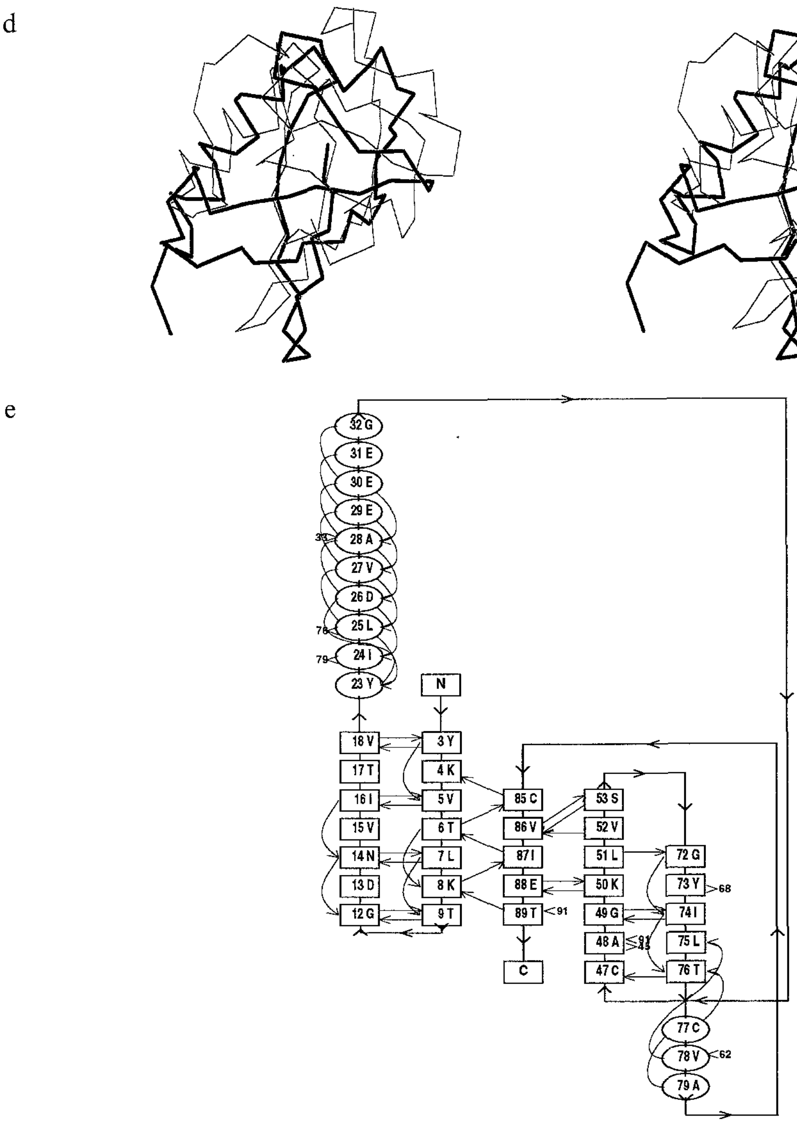

Fig. 5 (continued).

one location in the structure, the match would now be between a *set* of residues and the given location.

MODEL FOLD GENERATION

A fundamental limitation to all prediction methods based on fold recognition is that it is only possible to predict previously observed folds. We are tackling this problem by generating libraries of hypothetical model folds synthesized from observed folding patterns, and the currently known rules of protein folding. Two methods appear to be promising at present: polyhedral folding

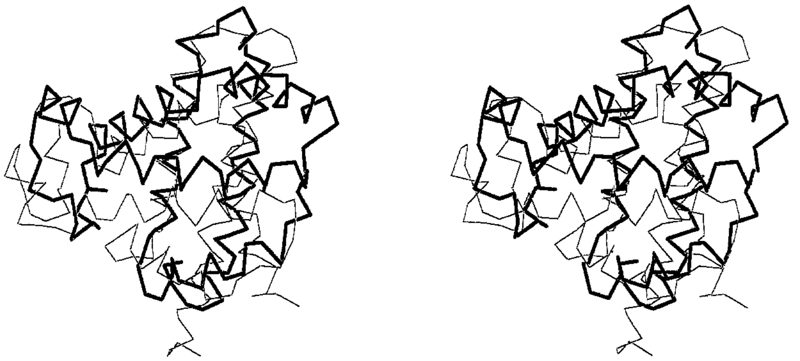

Fig. 6. A possible first step towards the creation of a library of hypothetical folds is illustrated here. The structure drawn with a thick line is the crystallographically determined structure for sperm whale myoglobin. The structure drawn with a thin line is a model for myoglobin, based on a regular polyhedral framework. The RMSD in this case is 3.6 Å over all 153 equivalent Cα atoms.

models and fragment build-up. Taylor [54] has shown that satisfactory models can be built for αα domains based on polyhedral frameworks, as proposed by Murzin and Finkelstein [55]. Current work suggests that a similar approach could be used for both αβ and ββ classes of protein structure. As an example of such a model, Fig. 6 shows the superposition of a model myoglobin structure with the crystallographically determined structure.

An alternative approach we are using is to analyse highly resolved proteins in order to derive a set of fragments representing known structural motifs. By generating novel combinations of these building blocks, using knowledge-based assembly rules, hypothetical model folds can be constructed. Simple filters may be applied to these hybrid models in order to eliminate models which violate protein folding rules, and which would therefore not be observed.

CONCLUSIONS

It is now well known that proteins such as the various TIM barrel enzymes, interleukin 1β/soybean trypsin inhibitor, and actin/hexokinase can show remarkable similarities in their native folds with no apparent sequence similarities. Furthermore, the rate at which newly solved protein structures are perceived to have previously observed folds suggests that the number of protein topologies may be limited. Indeed, some estimates put the number of observed topologies at 50% of the total number of naturally occurring topologies (although 10% is a more likely estimate). Given the significant possibility that a newly sequenced protein will have a previously observed fold, it is essential that methods for the recognition of protein folds in sequences be developed. Current success both in our own laboratory and in others is extremely encouraging, and it is to be hoped that these individual approaches to the problem will cross-fertilize and lead to even more successful methods in the future. A good example of this cross-fertilization is the method recently described by Godzik et al. [56] which incorporates aspects of both our optimal threading and the lattice-based approach of Finkelstein and Reva. Further developments can also be expected in the potentials used to evaluate the match of a sequence with a structure, and no doubt we can also expect more powerful computers to become available which will enable us to utilize ever more computationally intensive algorithms. With luck, by the time we have observed every possible fold, we will have the wherewithal to recognize these folds in our sequence data.

REFERENCES

1 Chou, P.Y. and Fasman, G.D., Biochemistry, 13 (1974) 212.

2 Garnier, J., Osguthorpe, D.J. and Robson, B., J. Mol. Biol., 120 (1978) 97.

3 Taylor, W.R., In Bishop, M.J. and Rawlings, C.J. (Eds.) Nucleic Acid and Protein Sequence Analysis; a Practical Approach, IRL Press, Oxford, 1987, pp. 359–385.

4 Fasman, G.D., Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989.

5 Nagano, K., J. Mol. Biol., 109 (1977) 251.

6 Nagano, K., J. Mol. Biol., 138 (1980) 797.

7 Taylor, W.R. and Thornton, J.M., Nature, 301 (1983) 540.

8 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

9 Taylor, W.R., J. Mol. Biol., 188 (1986a) 233.

10 Taylor, W.R., J. Theor. Biol., 119 (1986b) 205.

11 Pearl, L.H. and Taylor, W.R., Nature, 328 (1987) 351.

12 Bashford, D., Chothia, C. and Lesk, A.M., J. Mol. Biol., 196 (1987) 199.

13 Gribskov, M., Lüthy, R. and Eisenberg, D., Methods Enzymol., 188 (1990) 146.

14 Barton, G.J., Methods Enzymol., 188 (1990) 403.

15 Pastore, A. and Lesk, A.M., Proteins, 8 (1990) 133.

16 Rooman, M.J., Kocher, J.P.A. and Wodak, S.J., J. Mol. Biol., 221 (1991) 961.

17 Chan, H.S. and Dill, K.A., Proc. Natl. Acad. Sci. U.S.A., 87 (1990) 6388.

18 Doolittle, R.F., Protein Sci., 1 (1992) 191.

19 Flores, T.P., Orengo, C.A., Moss, D.S. and Thornton, J.M., Protein Sci., submitted.

20 Farber, G.K. and Petsko, G.A., Trends Biochem. Sci., 15 (1990) 228.

21 Chothia, C. and Finkelstein, A.V., Annu. Rev. Biochem., 59 (1990) 1007.

22 Orengo, C.A., Flores, T.P., Taylor, W.R. and Thornton, J.M., Protein Eng., in press.

23 Chothia, C., Nature, 357 (1992) 543.

24 Drexler, K.E., Proc. Natl. Acad. Sci. U.S.A., 78 (1981) 5275.

25 Novotny, J., Bruccoleri, R.E. and Karplus, M., J. Mol. Biol., 177 (1984) 787.

26 Brooks, B., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.

27 Novotny, J., Rashin, A.A. and Bruccoleri, R.E., Proteins, 4 (1988) 19.

28 Eisenberg, D. and McLachlan, A.D., Nature, 319 (1986) 199.

29 Baumann, G., Frommel, C. and Sander, C., Protein Eng., 2 (1989) 329.

30 Holm, L. and Sander, C., J. Mol. Biol., 225 (1992) 93.

31 Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 211 (1990) 959.

32 Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M.J., J. Mol. Biol., 216 (1990) 167.

33 Sippl, M.J., J. Mol. Biol., 213 (1990) 859.

34 Casari, G. and Sippl, M.J. (1983) J. Mol. Biol., 224 (1992) 725.

35 Crippen, G.M., Biochemistry, 30 (1991) 4232.

36 Maiorov, V.N. and Crippen, G.M., J. Mol. Biol., 227 (1992) 876.

37 Ponder, J.W. and Richards, F.M., J. Mol. Biol., 193 (1987) 775.

38 Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T., Proteins, 7 (1990) 257.

39 Bowie, J.U., Lüthy, R. and Eisenberg, D., Science, 253 (1991) 164.

40 Lüthy, R., Bowie, J.K. and Eisenberg, D., Nature, 356 (1992) 83.

41 Bork, P., Sander, C. and Valencia, A., Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 7290.

42 Finkelstein, A.V. and Reva, B.A., Nature, 351 (1991) 497.
43 Fauchere, J.L. and Pliska, V.E., Eur. J. Med. Chem., 18 (1983) 369.
44 Jones, D.T., Taylor, W.R. and Thornton, J.M., Nature, 358 (1992) 86.
45 Needleman, S.B. and Wunsch, C.D., J. Mol. Biol., 48 (1970) 443.
46 Taylor, W.R. and Orengo, C.A., J. Mol. Biol., 208 (1989) 1.
47 Kabsch, W. and Sander, C., Biopolymers, 22 (1983) 2577.
48 Kraulis, P.J., Science, 254 (1991) 581.
49 Vriend, G. and Sander, C., Proteins, 11 (1991) 52.
50 Benner, S.A. and Gerloff, D., Adv. Enz. Reg., 31 (1991) 121.
51 Thornton, J.M., Flores, T.P., Jones, D.T. and Swindells, M.B., Nature, 354 (1991) 105
52 Lesk, A.M. and Boswell, D.R., Bioessays, 14 (1992) 407.
53 Benner, S.A., Cohen, M.A. and Gerloff, D., Nature, 359 (1992) 781.
54 Taylor, W.R., Protein Eng., 4 (1991) 853.
55 Murzin, A.G. and Finkelstein, A.V., Biofizika, 28 (1983) 905.
56 Godzik, A., Kolinski, A. and Skolnick, J., J. Mol. Biol., 227 (1992) 227.
57 Hutchinson, E.G. and Thornton, J.M., Proteins, 8 (1990) 203.