Protein fold recognition by mapping predicted secondary structures

Robert B. Russell‡, Richard R. Copley and Geoffrey J. Barton†
University of Oxford
Laboratory of Molecular Biophysics
The Rex Richards Building, South Parks Road
Oxford, OX1 3QU, England
Tel: 44 1865 275368 FAX: 44 1865 510454
E-mail: gjb@bioch.ox.ac.uk
‡Present Address:
Biomolecular Modelling Laboratory
Imperial Cancer Research Fund Laboratories
44 Lincoln's Inn Fields, P.O. Box 123
London, WC2A 3PX, England
E-mail: russell@icrf.icnet.uk
†To whom correspondence should be addressed.
Keywords: protein; structure prediction; fold recognition; threading; nuclear magnetic resonance; secondary structure mapping.
Running title: Fold recognition from secondary structure prediction
Published in *J. Mol. Biol.* (1996), **259**, (3), 349-365.

Contents

1 abstract

A strategy is presented for protein fold recognition from secondary structure assignments (α -helix and β -strand). The method can detect similarities between protein folds in the absence of sequence similarity. *Secondary structure mapping* first identifies all possible matches (maps) between a query string of secondary structures and the secondary structures of protein domains of known three–dimensional structure. The maps are then passed through a series of structural filters to remove those that do not obey simple rules of protein structure. The surviving maps are ranked by scores from the alignment of predicted and experimental accessibilities. Searches made with secondary structure assignments for a test set of eleven fold-families put the correct sequence-dissimilar fold in the first rank 8/11 times. With cross-validated predictions of secondary structure this drops to 4/11 which compares favourably with the widely used THREADER program (1/11). The structural class is correctly predicted 10/11 times by the method in contrast to 5/11 for THREADER. The new technique obtains comparable accuracy in the alignment of amino acid residues and secondary structure elements. Searches are also performed with published secondary structure predictions for the von-Willebrand factor type A domain, the proteasome 20S α subunit and the phosphotyrosine interaction domain. These searches demonstrate how the method can find the correct fold for a protein from a carefully constructed secondary structure prediction, multiple sequence alignment and distance restraints. Scans with experimentally determined secondary structures and accessibility, recognise the correct fold with high alignment accuracy (86% on secondary structures). This suggests that the accuracy of mapping will improve alongside any improvements in the prediction of secondary structure or accessibility. Application to NMR structure determination is also discussed.

2 Introduction

The flood of new protein sequences demands techniques to infer protein 3D ‡ structure from sequence alone. For $\approx 30\%$ of protein sequences, conventional alignment techniques (e.g. (Lipman & Pearson, 1985; Altschul *et al.*, 1990; Smith & Waterman, 1981)) or profile and pattern methods (e.g. (Gribskov *et al.*, 1987; Barton & Sternberg, 1990)) find similarities to a protein of known 3D structure (Chothia, 1992). The remaining 70% of protein sequences may adopt previously unseen protein folds. Alternatively, they may have topologies (folds) similar to known protein structures but share no detectable sequence similarity (Russell & Barton, 1994). Such fold similarities will normally not be found until both protein 3D structures have been determined experimentally (Orengo, 1994; Holm & Sander, 1994*a*). In an attempt to find fold similarities of this type in advance of 3D structure determination, several fold recognition techniques have been developed. (see (Bowie & Eisenberg, 1993; Wodak & Rooman, 1993; Jones & Thornton, 1993) and refs therein.) These techniques may locate some fold similarities that are undetectable by the comparison of sequence. However, the methods are often computationally intensive and many similarities still go un-detected (Pickett *et al.*, 1992; Lemer *et al.*, 1996).

In parallel with the development of fold detection methods, the accuracy of secondary structure prediction has improved from $\approx 65\%$ to $\approx 72\%$ on average. Though this is only a small percentage increase, recent predictions are more useful, since the application of multiple sequence alignments improve the identification of the number, type and location of core secondary structure elements. Prediction from sequence alignments can also accurately identify the position of loops, and residues likely to be buried in the the protein core (Benner *et al.*, 1994; Barton, 1995; Russell & Sternberg, 1995). Given a good secondary structure prediction, the next question to ask is how the secondary structures might be arranged into a tertiary fold. *ab initio* methods for folding secondary into tertiary structure search for possible arrangements of secondary structures that obey general packing rules (Cohen & Sternberg, 1980; Cohen *et al.*, 1980; Cohen *et al.*, 1982; Smith-Brown *et al.*, 1993; Sun *et al.*, 1995). These methods have been applied in numerous blind predictions (Hurle *et al.*, 1987; Cohen *et al.*, 1986; Curtis *et al.*, 1991; Jin *et al.*, 1994; Huang *et al.*, 1994) with varied results. A limitation is the number of packing combinations that must be considered. This can become unmanageable for > 9 secondary structures (Cohen *et al.*, 1982), though approaches to reduce the number of combinations have been described (Taylor, 1991; Clark *et al.*, 1991).

The most successful predictions of protein tertiary structure in the absence of clear sequence similarity to a protein of known 3D structure, have been those where secondary structure predictions, and/or experimental information were combined to suggest resemblance to an already known fold. Correct folds have been predicted in this way for the α subunit of tryptophan synthase (Crawford *et al.*, 1987), a family of cytokines (Bazan, 1990), and recently, for the von Willebrand factor type A domain (Edwards & Perkins, 1995), and the synaptotagmin C2 domain (Gerloff *et al.*, 1995). Although the details of these studies differed, all used predicted secondary structures from multiple alignment, combined with the careful application of protein structural principles (often together with experimental data) to suggest a protein fold. Two automated methods for comparing predicted and experimental secondary structures have been described previously (Sheridan *et al.*, 1985; Rost, 1995) with promising though limited preliminary results.

In this paper we show how secondary structure and accessibility prediction together with basic rules of protein structure may be used to find the correct fold within a database of protein structural domains. The method first generates all possible matches (referred to as 'maps') between query and database secondary structure patterns, allowing for insertions and deletions of whole secondary structure elements. Maps are filtered by a series of structural criteria to arrive at a collection of sensible template structures. The sequence of the query protein is then aligned to the template structures by matching predicted and observed patterns of residue accessibility. Finally, alignments are ranked by a score that combines accessibility matching with a penalty for differences in secondary structure length. The method is designed to cope with incorrect secondary structure assignments, insertions/deletions of whole secondary structure elements, and differences in the lengths and orientations of secondary structures.

3 Methods

3.1 Database of unique protein 3D structural domains

A database of protein 3D structural domains was derived from the Brookhaven Protein Databank (Bernstein *et al.*, 1977). 930 non-identical chains were clustered by sequence comparison (Smith & Waterman, 1981; Barton, 1993) to leave 275 sequence families. One representative of each family was chosen to have the highest resolution and lowest R-factor. The representative structures were then split into 377 domains by eye. A sub–database of higher quality domains was created for analysis. This contained only those structures determined by X–ray crystallography, refined and of a resolution of 2.5 Å or better. Secondary structures for all domains were defined by the programs DSSP (definition of secondary structure in proteins) (Kabsch & Sander, 1983) or by DEFINE (Richards & Kundrot, 1988) when only C_α atoms were available. Axial coordinates were calculated for all secondary structures as described in (Richards & Kundrot, 1988). Extra axial coordinates were calculated at the N– and C– terminal ends to allow for possible differences in secondary structure length. The domain database is available via the WWW (http://geoff.biop.ox.ac.uk/).

3.2 Alignment of secondary structures

The secondary structure of the protein is represented as a sequence of H and B characters where each H represents an entire α helix and each B a β strand. A fast method for generating all exact matching alignments between two strings that allows up to a maximum number of deletions from each string (Russell *et al.*, 1995) is used to find all *maps* between the query pattern of secondary structures and the domain database. The method is recursive, and reminiscent of regular expression matching. In this study up to two deletions were permitted from the query secondary structure string, to allow for errors in the prediction. Up to five deletions were permitted from each database structure, to allow insertions or deletions of secondary structures typical of proteins having similar 3D structures in the absence of sequence similarity. Deletions from the database structure were only counted if they were contained within matched elements (overhanging deletions were ignored). Explicit mismatches were not allowed, but were treated as deletions from either the query or database structure. These values were chosen since they are typical of the expected accuracy of secondary structure prediction, and typical of insertions and deletions of secondary structure elements across members of a diverse structural family. In practice, the allowable deletions from query and database should be chosen on a case by case basis. For consistency, we kept the maximum numbers of deletions fixed during this study.

3.3 Filters

The alignment method will find all maps between two strings of secondary structure elements, but due to the allowance for deletions, many of these will correspond to implausible topologies. Accordingly, seven filters are used to remove maps corresponding to nonsensical protein 3D structures and/or those not satisfying imposed experimental restraints.

3.3.1 Removing un-compact structures

Two filters exploit the radius of gyration, R_g, to remove non-compact maps. Analysis of the 275 high quality domains shows that $R_g \leq 2.8L^{0.34} + 4.0$, where L is the length of the structure in residues. For each map, a coarse R_g is first calculated by considering the centroids of secondary structures, and their C-terminal loops as point masses. A fine R_g is also calculated by considering all matched residues (plus C-terminal loops) as point masses. Maps are removed if either R_g value is greater than the maximum for compact domains of the same length.

3.3.2 Loop length distance restraints

Analysis of the 275 high quality domains shows that the maximum distance D_{max} between axial coordinates that can be bridged by a loop of N_l residues is $11.621(N_l + 0.25)^{0.359} + 0.5$ Å. Maps having any loop with distances larger than $D_{max} + 4$ Å are removed. 4Å is added to allow for differences in the packing of database and query secondary structures, since similar structures with little sequence similarity can have shifts of up to 4 Å (Holm & C., 1995).

Care is taken to allow a range of possible positions for the match of query and database structures. This allows for errors in secondary structure prediction, which may fail to predict the precise start or end of correctly identified elements, and allows for the observed differences between the lengths of secondary structure elements within proteins having similar topologies despite no significant sequence similarity. For a position x on a database secondary structure, and a minimum and maximum length for a query secondary structure, L_{min} and L_{max}, the range of allowable positions of the query residue on the database structure (of length L_{obs}) is given by:

$$x_{min} = \min\left(L_{obs} - L_{max}, 0\right) - h + x$$
$$x_{max} = \max\left(L_{obs} - L_{min}, 0\right) + h + x$$

where h is a leniency parameter, allowing for differences in the length of query and database secondary structures. $h = 4$ allows for differences typical of those found in proteins having similar 3D structures despite no sequence similarity.

3.3.3 Poor β sheets

The deletion of β strands from a β sheet can lead to maps corresponding to nonsensical 3D structures. Maps containing isolated β strands (i.e. those lacking hydrogen bonding partners) are removed. Maps are also removed if β strands are deleted from the centre of β sheets contained within the map.

Analysis of high quality domains shows that the number of C_α – C_α contacts ≤ 6 Å made by a β strand ($C_{\beta-\alpha\alpha}$) with any of its neighbouring β strands is always $\geq N_\beta - 2$, where N_β is the number of residues in the β strand. Thus maps are also removed if one or more β strands has $C_{\beta-\alpha\alpha} < N_\beta - 2$.

3.3.4 Adjacent parallel structures

Maps are removed if tandem secondary structures in the query are made to match parallel structures in the database by the deletion of intervening secondary structures. Genuine adjacent parallel structures within the database are allowed. This filter can be turned off in instances when there are long loops connecting query secondary structure elements, as in the phosphotyrosine interaction domain example (see Results).

3.3.5 Distance restraints

Distance restraints may be imposed from the results of NMR experiments, knowledge of the disulphide linkages, or knowledge of residues involved in the active or binding site of the query. In this study, distance restraints are only included in the von Willebrand factor and Proteasome examples (see results). A tolerance value $t = 4\text{Å}$ is added to all distance restraints as for the loop length filtering.

3.3.6 Consistency & Redundancy

Maps are only kept if there is at least one placement of the query onto the database secondary structures where all distance restraints (loop length and/or experimental) are satisfied simultaneously.

After application of all the other filters, matches contained entirely within another match are considered redundant, and removed.

3.3.7 Maps removed by each filter

It is illustrative to consider the fraction of maps removed by each of the filters described above. For example, a pattern derived from a DSSP assignment of secondary structure for thioredoxin that allows for 2 secondary structure element deletions from the query and 5 from the database, the initial alignment of secondary structure elements reduces the number of folds from 377→212. 165 folds have no match of secondary structures with the predicted thioredoxin pattern. Table 1[1] illustrates the fractions of the initial 204783 maps within 212 folds that are removed by each filter when applied independently. Table 2[2] shows for the same example, how the number of maps drops as the filters are applied in succession. The filters are independent of one another apart from consistency filtering, which must be applied *after* loop and distance restraint filtering, and redundancy filtering, which must be applied last. The order of filters shown in Table 1[3] was chosen so as to optomise speed.

The gradual elimination of maps and folds shows how the simple principles

[1] ntable1.ps
[2] ntable2.ps
[3] ntable1.ps

of protein structure are sufficient to reduce the number of possible alignments by two orders of magnitude. Interestingly, the number of folds drops very little after the generation of maps, suggesting that the filters are tending mostly to remove nonsensical maps associated with each identified fold rather than ruling out folds. Note that consistency filtering tends only to remove maps when tight loop lengths or distance restraints are included in the pattern.

3.4 Fitting sequences on to 3D structures

Accessibilities for residues within each map are calculated quickly by explotting the relationship between relative accessibility and the number of other C_β atoms within 7 Å ($N_{C\beta 7}$) of a residues C_β atom. $N_{C\beta 7}$ is calculated by considering secondary structures and the C-terminal coils for the matched structures. Analysis of the high quality domains shows that helical residues are buried (b) when $N_{C\beta 7} \geq 3$, exposed (e) when $N_{C\beta 7} = 0$ and intermediate/unknown (u) otherwise. Similarly, residues in β strands are b when $N_{C\beta 7} \geq 6$, e when $N_{C\beta 7} \leq 3$ and u otherwise. In the examples presented here, predicted accessibilities were taken from the SUB line within PHD (Rost & Sander, 1994) output, which highlights those regions predicted with confidence. Remaining positions were assigned as unknown (u) accessibility.

Given assignments of accessibility, the best alignment for each pair of secondary structures not permitting gaps within either secondary structure is found by applying the scoring matrix shown in Table 3[4]. These values were chosen to prevent long overhanging gaps in the alignment of predicted and experimental secondary structures, and designed not to penalise mismatches too heavily. The total similarity score for the alignment is then defined as:

$$(\sum_{i=0}^{i=N} S_{acc}) - L_{diff}$$

where S_{acc} is the best score for a pair of matched secondary structures calculated by summing values from Table 3[5], N is the number of matched secondary structures, and L_{diff} is the total difference in the lengths of the two

[4]ntable3.ps
[5]ntable3.ps

protein domains being compared. When calculating L_{diff} those secondary structures that have been equivalenced are ignored, since overhanging gaps are already penalised by the gap score in Table 3[6].

3.5 Protein Structure Patterns for Evaluation

Representatives (queries) from each of 11 structural families containing structural similarities despite no sequence similarity (Russell & Barton, 1994) were chosen to assess the method. The 11 queries are shown in Table 4[7] and represent a diversity of folds from all four protein folding classes. For all queries, there is at least one clear example of a similar fold in the database that does not show any detectable sequence similarity to the query. For reference, similar folds in the database were found by the STAMP (structural alignment of multiple proteins) structure comparison program (Russell & Barton, 1992) and with reference to the structural classification of proteins (SCOP) database (Murzin *et al.*, 1995).

Two patterns were defined for each of the eleven structures: a) one taken directly from the DSSP secondary structure assignment and accessibility (i.e. perfect prediction) and b) one from cross–validated secondary structure and accessibility prediction by the methods of Rost & Sander (Rost & Sander, 1993; Rost & Sander, 1994). The PHD program and jack-knifed neural network architectures were kindly provided by Dr Burkhard Rost (EMBL). Experimental secondary structure summaries and accessibilities (a) were taken from DSSP (Kabsch & Sander, 1983). Predicted secondary structure summaries (b) were taken from the 'PHD sec' entries and accessibilities from the 'SUB acc' entries, since these most closely resembled the assignments from the $N_{C\beta 7}$ calculation of accessibility. PHD assignments of buried and exposed states were classified as buried and exposed, with all other positions 'i' or no assignment as 'u'. Strands shorter than two residues, and helices shorter than four residues were ignored. The length of the secondary structure was given by the number of residues in each secondary structure (maximum = minimum), and the number of residues between the secondary structures was taken as the minimum loop length.

[6] ntable3.ps
[7] ntable4.ps

Patterns may also contain distance restraints, such as those available from NMR experiments, disulphide linkages, or SDM studies. Distance restraints were only added in the von–Willebrand factor and Proteasome patterns (see Results).

3.6 Cross–validation

Any predictive method that needs large numbers of parameters must be cross-validated to ensure that the method does not do artificially well on the examples used to derive the parameters. For cross validation of the secondary structure and accessibility predictions, we used the jack-knifed neural–network architectures described by Rost & Sander (1993a) (Kindly provided by Dr. B. Rost.) Secondary structure and accessibility for each query protein was predicted by an architecture that did not include the query protein or any homologue.

The filters and matching algorithm described here use only a few geometric parameters all of which are independent of the protein sequence. Accordingly, removal of query proteins and homologues from the set used to derive the equations above makes a negligible difference to the parameters.

3.7 Computational details

Runs for the patterns shown in Table 4[8] take between 5 and 60 minutes on a Silicon Graphics Indigo 2 (150 MHZ IP22 Processor MIPS R4400). The MAP program is available from the authors. Contact GJB by e-mail: gjb@bioch.ox.ac.uk or see the WWW address http://geoff.biop.ox.ac.uk/ for details.

[8] ntable4.ps

4 Results

4.1 Assessing accuracy

Structural similarity is a continuum and for some fold types opinions differ as to what constitutes "similar". For example, thioredoxin has a β-sheet with helices packing on each side which superficially resembles a Rossmann fold domain. However, the topology of the sheet is different from a Rossmann fold: the connectivity is different, and it contains a mixture of parallel and antiparallel β hairpins rather than all parallel. To build a detailed model of thioredoxin based on a Rossmann fold would be incorrect, but recognising that thioredoxin has a "single sheet with helix on each side" is still useful. For some folds, e.g. the β-trefoils, there is no such ambiguity. We discuss the accuracy of our method using two grades of success 'strict' and 'loose', which are outlined in Table 5[9]. Strict similarities are those where the topology of the structure in the database is nearly an exact match of that found in the query (e.g. plastocyanin and azurin). Loose similarities are those where the topologies are broadly similar, with additional secondary structures in one fold relative to another, and with some differences in topological ordering or orientation of equivalent secondary structure elements (e.g. plastocyanin and an Ig fold). Strict similarities tend to correspond with those specified by **scop** (Murzin *et al.*, 1995), whereas the loose similarities tend to correspond roughly with those identified by CATH (Orengo *et al.*, 1993) and by the assessors of the protein structure prediction challenge (Lemer *et al.*, 1996).

For comparison, we also scanned the same eleven queries against the database of domains using the fold recognition program THREADER (Jones *et al.*, 1992) with default parameters.
In addition to the recognition of the correct fold, it is important to consider how well the query is aligned onto the database structure. Two measures of alignment accuracy are given: a) the fraction of correct residue equivalences found by each method *% Res–Res*, and b) the fraction of correctly overlapping secondary structure elements found *% Sec–Sec*. Secondary structures were considered correctly matched if at least two residues from structurally equivalent secondary structures overlapped in the alignment generated by

[9]ntable5.ps

each method. % Res–Res is a strict definition, and broadly measures how accurate a 3D model would be if based on the alignment found. % Sec–Sec is a looser definition, and allows for slippages of secondary structures and thus indicates the accuracy of the predicted topology. The second measure is arguably a more reliable guide, since for many pairs of similar protein structures, alignments of sequence based on 3D structure are ambiguous. Problems arise when assessing the symmetrical α/β barrel structures. Shifting the alignment of secondary structure elements by one β α unit can lead to zero accuracy by these measures, though the resulting structure is largely correct. We thus report average accuracies with and without the α/β barrels. To assess the overall alignment accuracies of each method, only those strict similarities that were not detectable by a sensitive sequence comparison algorithm (Barton, 1993) were considered. Similarities excluded were those with the globins, 1ECA, 1HBG and 1MYGA when scanning with Sea Hare Myoglobin, and that with 1PAZ when scanning with plastocyanin. For all other examples, accuracies were included in the calculation of an average, regardless of whether the similarity was found at or near the top of the ranked lists. A total of 36 strict similarities were used in the calculation.

4.2 Searches with eleven test proteins

The results of comparing the eleven protein structures to the database of domains using DSSP patterns, PHD patterns, and the THREADER program are shown in Table 6[10]. The table lists the top 10 ranked domains for each query by each method. For each domain, the code, score, structural class and fold description are shown together with the alignment score and the percentage accuracies of the alignments at the residue (% Res-Res) and secondary structure (% Sec–Sec) level (see below). Within Table 6[11], domains classified as strict similarities (ignoring those detectable by sequence comparison) are shown in inverse text; loose similarities are shown as shaded. Table 7[12] summarises the rankings shown in Table 6[13] (see legend).

[10]ntable6.ps
[11]ntable6.ps
[12]ntable7.ps
[13]ntable6.ps

Judging by the strict criteria shown in Table 5[14], 8/11 of the scans made with experimentally determined secondary structure (MAP(DSSP)) put the correct fold in the first rank. By the loose definition, the method located 10/11 folds in the first rank. Predictably, the scans based on patterns from secondary structure prediction fare worse. 4/11 folds were correctly ranked at position 1 by the strict criteria. However, this compares favourably with THREADER which placed 1 fold correctly in the first rank. When the loose definitions of fold similarity are used, our method placed 5/11 correct folds at the top of the list compared to 2/11 for THREADER. Expanding the definition of success to include any search that places a correct fold in the top 10, as described by Lemer *et al.* (1996) , shows a similar trend (Table 7)[15]. The greater success of the DSSP derived patterns suggests that fold recognition by this method will improve alongside any improvements in secondary structure and accessibility prediction. The structural class of proteins (as identified using SCOP) in the top 10 domains was more consistent by our method: MAP(PHD) scans lead to 10/11 correct protein class predictions for the 1st ranked protein, compared to 5/11 for THREADER. Although this improvement may be due mostly to the accuracy of the PHD predictions, the result suggests that other fold recognition methods could profit from the consideration of predicted secondary structures.

Our method (MAP) shows an improvement over THREADER with respect to detecting the correct fold. What of alignments of sequence to structure? Values for individual accuracies are given in Table 6[16]. Reference alignments of 3D structures were found by the STAMP algorithm (Russell & Barton, 1992) for all strict similarities with the eleven protein families. The averaged values for % Res–Res and % Sec–Sec are shown in Table 8[17]. MAP(DSSP), MAP(PHD) and THREADER give % Res–Res of 35, 15 and 11 % respectively and % Sec–Sec of 75, 43 and 37%. If one ignores the repetitive α/β barrel alignments, accuracies improve slightly with % Res–Res 39, 15 and 13% and % Sec–Sec of 86, 49 and 50 % for MAP(DSSP), MAP(PHD) and THREADER. None of the methods perform well by the % Res–Res criterion, though % Sec–Sec suggests that the correct topology

[14]ntable5.ps
[15]ntable7.ps
[16]ntable6.ps
[17]ntable8.ps

is achieved about 50 % of the time. The high % Sec–Sec for MAP(DSSP) scans suggests that alignment accuracy, like fold recognition, will improve with developments in secondary structure and accessibility prediction.

How useful are the detected loose similarities? For some examples, loose similarities imply only a broadly similar architecture, and may not immediately be used for homology modelling studies. However, for others the loose similarity genuinely represents a feasible modelling template. For example, the PHD prediction of hepatocyte nuclear factor 3 (HNF-3) failed to predict two short β strands found in the native structure, and thus the MAP search did not detect BirA domain I (PDB code 1BIA) or GAP domain I (2GAP) as possible templates. However, the search with the predominantly helical prediction did rank another helix-turn-helix motif first, as shown in Figure 1[18]. The core three helices have been aligned correctly at the secondary structure level and a prediction of this type could be useful in the absence of experimental 3D structure information.

4.3 Fold recognition from published predictions

In the tests above only the type and length of secondary structures, the loop length observed in the query structure, and the pattern of burial and exposure, observed or predicted for each secondary structure segment were used in the search. Many published predictions are augmented by human insight, contain detailed predictions of loop lengths, and consider experimental distance restraints. All of this information can be used with the MAP method described here. To test the method under these circumstances, we considered three predictions: 1) the von Willebrand factor (vWf) prediction by Edwards & Perkins (1995), 2) the Proteasome prediction by Lupas *et al.* (1994) and 3) a prediction for the Phosphotyrosine Interaction Domain (PID) by Bork & Margolis (1995). All of these predictions were made from very diverse sequences, which is likely to improve prediction accuracy (Russell & Sternberg, 1995). The predictions also comprise carefully constructed sequence alignments, that can provide tight loop–length distance restraints. For the three searches, a larger and more up-to-date database of 780 protein domains was scanned (A. S. Siddiqui per. comm.) Subsequent 3D structure determination

[18]figure1.ps

has shown all three of these proteins to resemble previously observed folds (Lee *et al.*, 1995; Brannigan *et al.*, 1995; Zhou *et al.*, 1995).

4.3.1 The vWF domain

Perkins & co–workers (Perkins *et al.* , 1994; Edwards & Perkins, 1995) used an alignment of 92 sequences together with spectroscopic data, and prediction algorithms to predict that the vWf domain would comprise a repeating arrangement of β strands and α helices. Edwards & Perkins combined a THREADER scan with analysis of the location of active site residues, a putative disulphide bridge, and the principles of protein 3D structure. They suggested that the vWf domain would be most likely to resemble ras p21. The subsequently determined 3D structures (Lee *et al.*, 1995) showed this prediction of secondary structure and fold to be largely correct (Russell & Sternberg, 1995).

Our mapping technique allows many of the features exploited by Perkins *et al.* to be combined in a prediction. Figure 2[19] shows a vWf pattern based on the prediction of Perkins & co-workers (Perkins *et al.*, 1994; Edwards & Perkins, 1995). In addition to a pattern of predicted secondary structures, the pattern also contains detailed information as to the loop lengths, and details of two distance restraints: one from a pair of aspartic acids thought to be involved in a metal binding site (constrained to have their axial coordinates within 15 Å), and a putative disulphide bond (constrained to have their axial coordinates within 9.5 Å). A tolerance of $t = 4$ Å was added to each of these restraints to allow for changes in secondary structure packing across similar protein 3D structures.

A comparison of the vWf pattern to the database of 780 domains finds Elongation factor Tu (PDB code 1ETU), Ras P21 (821P) and Che-Y (3CHY) as the three top scoring folds, with other double–wound, α/β , Rossmann-type folds following in the top 20 scoring folds. The top 3 scoring proteins are highly similar to the recently solved structures of the vWf, with Ras P21/Elongation factor Tu being the most similar (Lee *et al.*, 1995).

[19] figure2.ps

4.3.2 The Proteasome

Lupas *et al.* (1994) predicted the secondary structure for the 20S proteasome α subunits by a variety of algorithms. We took their predicted pattern of secondary structure elements and accessibility and searched the database of 780 non-redundant protein domains. Without imposing any experimental distance restraints, the method finds 7 folds (173 maps). The top scoring fold, according the the amphipathicity scoring scheme, is that of glutamine amidotransferase (PDB code 1GPH), which is structurally and functionally similar to the proteasome (Lowe *et al.*, 1995; Brannigan *et al.*, 1995).

A small number of weak distance restraints can make a significant difference to the results of this search. If alignment positions identified as putative active site residues by Lupas *et al.* , by the method of Benner and co-workers (Benner *et al.*, 1993), are required to have axial coordinates within 15 Å (tolerance of 4 Å) of each other, only 4 folds (19 maps) remain, with the correct fold still at the first rank. Although distance restraints are not always available prior to 3D structure determination, our results suggest that they should be used to aid fold recognition whenever possible.

4.3.3 The phosphotyrosine interaction domain

Bork & Margolis (1995) recently identified a new phosphotyrosine interaction domain (PID) involved in the cytoplasmic signalling cascade. They constructed an alignment of several diverse members of this sequence family, and performed a prediction of secondary structure. We ran the PHD program on a slightly more up-to-date alignment of PID proteins (P. Bork, personal communication), to predict the secondary structure and accessibility. A search pattern was made from the prediction, and the loop length ranges taken from the multiple alignment. The pattern of 9 secondary structures was BBHBBBBBH and these elements are numbered sequentially from 1–9 below. Since there were two long loops connecting the predicted secondary structures, the adjacent parallel filter was not used during the search. Structures corresponding to the best alignment with each of the top six scoring folds are shown in Figure 3[20]. Recent structure determination has shown the PID (PTB domain) to resemble the plekstrin homology (PH) domain

[20]figure3.ps

in structure and function (Zhou *et al.*, 1995). By the accessibility scoring scheme, the top ranked fold is not a PH domain, although a PH domain (from dynamin) is ranked at position 2. The top 6 folds are illustrative in that they show how the method can suggest alternative plausible folds that satisfy a pattern of predicted secondary structures and accessibilities.

The best scoring fold (Figure 3a)[21] is that of profilin (PDB code 2BFPP), and the best scoring map gives an anti-parallel β sheet with the strand order 218754 (predicted strand 6 is deleted) with one helix packing against each face. The second best scoring fold is a correct match with the PH domain from human dynamin (1DYNB), having deleted the first predicted α helix from the PID pattern. The third best scoring fold (3c) comes from *S. aureus* β lactamase (1BLH, domain 1), with an anti-parallel β sheet of order 54876, with both helices packing against one face. The fourth and fifth best scoring folds come from members of the Ig superfamily, and comprise alternative arrangements of β strands to form a greek key β sandwich. Both of the predicted α helices from the PID pattern have been deleted in these matches. Finally, the sixth (3e) match comes from the tryptic core of *E. coli lac* repressor (1TLFD domain 4), and comprises a parallel β sheet (42576) with both helices packing against one face. This fold is perhaps the least plausible, since it would require 3 crossover connections between adjacent and parallel β strands.

The method has suggested plausible alternative structures that can be scrutinised, in the absence of 3D structural data, by way of further experiments, secondary structure predictions, or even other methods of fold recognition. The results show how the predicted secondary structure elements can be accommodated into a compact, plausible protein fold, and encouragingly, the method has identified the correct fold high in the list of alternatives.

5 Discussion & Conclusions

In this paper we have presented a new method for protein fold recognition which exploits recent improvements in protein secondary structure prediction, and can use other information such as predictions of accessibility, loop

[21] figure3.ps

lengths and experimental data to restrict possible folds. When applied to predicted secondary structures and accessibilities, the method has been shown to be slightly better than one widely used fold recognition method (Jones *et al.*, 1992) at detecting the correct fold for eleven test examples. The alignments generated by the method are of comparable accuracy at the residue-residue and secondary structure alignment level. When the query is defined by experimental secondary structures and accessibilities, the method is highly successful at recognising the correct fold. This suggests that the mapping method will improve alongside any future improvement in secondary structure and accessibility prediction. The method also has the advantage of being computationally inexpensive, and so allows for multiple searches to be performed quickly.

The simplicity of the technique suggests several enhancements that could improve accuracy even further. The method of aligning sequences onto 3D structures might be developed by the use of empirically derived pair-potentials or accessibility preferences (e.g. (Sippl, 1990; Jones *et al.*, 1992)), or by the identification of favourable interaction sites between secondary structures (Cohen & Sternberg, 1980; Cohen *et al.*, 1980; Cohen *et al.*, 1982). A more sophisticated alignment and ranking procedure is under development.

The initial alignment and filtering procedures are perhaps the most unique feature of this technique. Other techniques for fold-recognition tend only to provide a single sequence alignment of query and database structures. The use of a secondary structure element alignment method has the advantage that exhaustive comparisons of two proteins can be performed; most folds identified have an ensemble of alternative alignments that can be explored further.

Since most protein structure similarities occur at the domain level, it is advantageous, whenever possible to split both query and database structures into domains. The problem of assigning domains for protein 3D structures has been the subject of revived interest (Holm & Sander, 1994*b*; Siddiqui & Barton, 1995; Sowdhamini & Blundell, 1995; Islam *et al.*, 1995) and is likely to lead to accessible databases of protein structural domains. Assigning domains within proteins of unknown 3D structure is more problematic, though approaches based in sequence homology (Pongor *et al.*, 1994; Sonnhammer &

Kahn, 1994) are undoubtedly the most promising; the vWf and PID proteins above are both examples of domains that occur in a variety of multi-domain contexts.

The method described here has applications in protein structure determination by NMR. During NMR structure determination, a preliminary secondary structure assignment (equivalent to a very accurate prediction) and a small number of distance restraints may be available early in the study. However, these data are usually insufficient to determine a unique structure by distance geometry or molecular dynamics (Smith-Brown *et al.*, 1993). Our results for the vWF and Proteasome domains suggest that the data may be sufficient to locate a similar fold in the database if one is present. Folds predicted from distance restraints and secondary structure assignment may be used to guide the assignment of cross-peaks and thus speed up the structure determination process. Clearly, the alternative consistent topologies may also give clues as to possible structural/functional/evolutionary relationships that are generally not known until after 3D structure determination (such as that described in Matthews *et al.*, 1994).

We have shown that secondary structure predictions of typical accuracy, together with simple principles of protein 3D structures and/or experimental data can be used to recognise correct protein folds in a library of domains. These results and others (Edwards & Perkins, 1995; Russell & Sternberg, 1995; Gerloff *et al.*, 1995) suggest that secondary structure prediction, experimental data, and protein structural principles should be used to augment protein fold recognition whenever possible.

6 Acknowledgements

We thank Professor L.N. Johnson for encouragement and support. We are indebted to Dr D.T. Jones (University of Warwick) for giving advice on the THREADER program and its database, Dr B. Rost (EMBL, Heidelberg) for providing the PHD program, Drs P. Bork (EMBL, Heidelberg) and S.J. Perkins (Royal Free Hospital, London) for providing prediction data for the PID and vWF domains, Dr S.K. Burley (Rockefeller University, New York) for providing the coordinates of the HNF–3 structure and Mr A. S. Siddiqui

(LMB, Oxford) for providing a database of protein structural domains. RBR thanks Dr C. P. Ponting (Fibrinolysis Research Unit, Oxford) for helpful discussions. RBR and GJB thank the Royal Commission for the Exhibition of 1851 and the Royal Society for support. RRC is funded by an MRC studentship. This research was funded in part by a grant from the BBSRC (UK).

7 ‡Abbreviations

3D three dimensional; NMR nuclear magnetic resonance; Ig Immunoglobulin; SDM site directed mutagenesis; WWW world wide web; The standard one– and three–letter abbreviations for the amino acids are also used throughout.

8 Figure and Table Legends

8.1 Figure 1

Figure 1[22].

An example of a useful 'loose' similarity between 3D structure detected using the MAP method and a secondary structure prediction. a) The alignment found by the method between the predicted pattern for HNF–3 and the helical DNA binding motif within phage 434 repressor. Boxed, bold-faced, upper-case regions indicate aligned predicted and experimental secondary structures. Sec denotes the PHD prediction for HNF–3, and a 3-state DSSP secondary structure assignment for 434 repressor. Bur shows predicted and experimental states of burial for HNF–3 and 434 repressor: b = buried, e = exposed; u = intermediate/unknown. b) The equivalent alignment found using the STAMP (Russell & Barton, 1992) structure comparison algorithm. Boxed, bold-faced, upper-case regions indicate structural equivalences. Sec denotes DSSP 3–state secondary structures for both proteins. c) and d) show the crystallographic structures of the matched regions of HNF–3 and 434 repressor, with structurally equivalent residues shown in ribbon/coil format, and unequivalent regions shown as C_α trace. The N- and C- termini of the

[22]figure1.ps

structures are labelled.

8.2 Figure 2

Figure 2[23].

Search pattern for the von-Willebrand factor type A domain (derived from Edwards & Perkins, 1995) as discussed in the text. α helices are indicated by cylinders, β strands by arrows. The range of numbers given beside each secondary structure or loop are the range of predicted lengths. Bullets ($\bullet$) show those secondary structure that are required for any possible map (i.e. those involved in distance restraints). Two distance restraints, one from a putative disulphide bond (9.5 Å)and the other from knowledge of two residues thought to be involve in metal coordination (15 Å) are shown to the left of the figure.

8.3 Figure 3

Figure 3[24].

Maps from the top six scoring folds found during a search with the PID pattern. Details are given in the text.

8.4 Table 6

Table 6[25].

Results of running MAP using secondary structure assignments (I) and PHD secondary structure predictions (II) shown beside THREADER results (III) for eleven protein structures having type B and C similarities (Russell & Barton 1994) within the domain database. The first column for each method shows the top ten scoring domains, which are denoted by a PDB four letter code (Bernstein *et al.*, 1977), a chain identifier as the fifth character (if any), followed by an underscore and a Roman numeral denoting

[23]figure2.ps
[24]figure3.ps
[25]ntable6.ps

the domain (if any). Bold inverted text denotes a correct match using the strict classification, grey backgrounds show loose classifications (see text). The second column shows the score for each domain, the third the protein structure class, and the fourth the name of the fold/structure. Upper case denotes fold families under the strict definitions. Upper case names in parentheses (if present) denote the name of the loose family classification. The globins 1HBG, 1MYGA and 1ECA and the cupredoxin 1PAZ are sequence similar to the query so are not shown inverted and are not included in the evaluation statistics (see text).

Strict fold classifications: 4HB-1= Up-down-up-down four helix bundle (4HB); 4HB-2= up-up-down-down (interleukin-4 type) 4HB; GLOBIN= globin-type folds; W-HTH= winged helix-turn-helix (HTH) folds; EF-HAND= calcium binding EF hands; CYTOC= cytochromes C; THIO= thioredoxin-like folds; FLAVO= flavodoxin-like folds; ROSS= Rossman folds; PBL= periplasmic binding protein-like folds; ACTIN-ATPASE= actin/HSC-70/hexokinse like folds; G-PROT= G-protein (ras) like folds; FAD-BIND= FAD/NAD binding protein-like folds; α β -BARREL= α β (TIM) barrels; β -GRASP= β -grasp (ferredoxin) like folds; IG= Immunoglobulin superfamily; CUP= Cuppredoxins (plastocyanin-like); β -TREFOIL= β -trefoils (interleukin-1-β -like); OB-FOLD= oligonucleotide/oligosaccharride binding folds.

Loose fold classifications: 4HB= 4HB-1, 4HB-2, ferritin; HTH= W-HTH, λ-rep., trp-rep.; DWAB (doubly-wound-α β)= ROSS, FLAVO, THIO, PBL, G-PROT, sugar phosphatase, pfk, pgk, dhfr; GKBS (greek key β sandwich)= IG, CUP, α -amylase inhibitor, sod, macromycin, prealbumin.

Other abbreviations: sod: superoxide dismutase; pfk= phosphrofuctokinase; pgk= phosphoglcerate kinase; dhfr= dihdrofolate reducatse; ldh= lactate dehydrogenase; ser-prot= serine proteinase; asp-prot= aspartic proteinase; inh.= inhibitor; rep.=repressor; glut.=glutathione; red.=reductase; thym. phosph.=thymidine phosphorylase; ribo.=ribonuclease; glyc.=glycoprotein; P-glucomutase= phosphoglucomutase; glyc. ribo trans.= glycinamide ribotransferase.

8.5 Table 7

Table 7[26].

[26] ntable7.ps

Summary of fold recognition success rates. Strict and Loose refer to the critera for structural similarity discussed in the text. Class refers to structural class success as discussed in the text. (1st) refers to success measured as a correct fold at rank 1, (Top 10) as a correct fold in the top 10 ranked structures.

9 Figures

[Figure 1[27]][Figure 2[28]][Figure 3[29]].

10 Tables

[Table 1[30]][Table 2[31]][Table 3[32]][Table 4[33]][Table 5[34]][Table 6[35]][Table 7[36]][Table 8[37]].

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Barton, G. J. (1993). An efficient algorithm to locate all locally optimal alignments. *Comp. App. Biosci.* **9**, 729–734.

Barton, G. J. (1995). Protein secondary structure prediction. *Current Opinion Structural Biol.* **5**, 372–376.

[27] figure1.ps
[28] figure2.ps
[29] figure3.ps
[30] ntable1.ps
[31] ntable2.ps
[32] ntable3.ps
[33] ntable4.ps
[34] ntable5.ps
[35] ntable6.ps
[36] ntable7.ps
[37] ntable8.ps

Barton, G. J. & Sternberg, M. J. E. (1990). Flexible protein sequence patterns - a sensitive method to detect weak structural similarities. *J. Mol. Biol.* **212**, 389–402.

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci.* **87**, 6934–6938.

Benner, S. A., Badcoe, I., Cohen, M. A. & Gerloff, D. L. (1993). Predicted secondary structure for the *src* homology 3 domain. *J. Mol. Biol.* **229**, 295–305.

Benner, S. A., Gerloff, D. L. & Jenny, T. F. (1994). Predicting protein crystal structures. *Science,* **265**, 1642–1644.

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

Bork, P. Margolis, B. (1995). A phosphotyrosine interaction domain. *Cell,* **80**, 693–694.

Bowie, J. U. & Eisenberg, D. (1993). Inverted protein structure prediction. *Current Opinion Structural Biol.* **3**, 437–444.

Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C. E., Smith, J. L., Tomchick, D. R. & Murzin, A. G. (1995). A protein catalytic framework with and N–terminal nucleophile is capable of self-activation. *Nature,* **378**, 416–419.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature,* **357**, 543–544.

Clark, D. A., Shirazi, J. & Rawlings, C. J. (1991). Protein topology prediction through constraint-based search and the evaluation of topological folding rules. *Prot. Eng.* **4**, 751–760.

Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L. & Smith, K. A. (1986). Structure–activity studies of interleukin–2. *Science,* **234**, 349–352.

Cohen, F. E. & Sternberg, M. J. E. (1980). On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. *J. Mol. Biol.* **137**, 9–22.

Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1980). Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature,* **285**, 378–382.

Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. (1982). The analysis and prediction of the tertiary structure of globular proteins involving the packing of α helices against a β sheet: a combinatorial appraoch. *J. Mol. Biol.* **156**, 821–862.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Predictions of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. *Proteins,* **1**, 118–129.

Curtis, B. M., Presnell, S. R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffery, E., Cosman, D., March, C. J. & Cohen, F. E. (1991). Experimental and theoretical studies of the three-dimensional structure of human interleukin–4. *Proteins,* **11**, 111–119.

Edwards, Y. J. K. & Perkins, S. J. (1995). The protein fold of the von willebrand factor type A is predicted to be similar to the open twisted β-sheet flanked by α-helices found in human ras-p21. *FEBS Letters,* **358**, 283–286.

Gerloff, D. L., Chelvanayagam, G. & Benner, S. A. (1995). A predicted consensus structure for the protein-kinase c2 homology (c2h) domain, the repeating unit of synaptotagmin. *Proteins,* **22**, 299–310.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Nat. Acad. Sci.* **84**, 4355–4358.

Holm, L. & C., S. (1995). 3-d lookup: fast protein structure database searches at 90 % reliability. *Proc. 3rd Int. Conf. Intel. Sys. Mol. Biol.* , 179–187.

Holm, L. & Sander, C. (1994*a*). Searching protein structure databases has come of age. *Proteins,* **19**, 165–173.

Holm, L. & Sander, C. (1994*b*). Parser for protein folding units. *Proteins,* **19**, 256–268.

Huang, Z., Gabriel, J., Baldwin, M. A., Fletterick, R. J., Prusiner, S. B. & Cohen, F. E. (1994). Proposed three–dimensional structure for the cellular prion protein. *Proc. Nat. Acad. Sci.* **91**, 7139–7143.

Hurle, M. R., Matthews, C. R., Cohen, F. E., Kuntz, I. D., Toumadje, A. & Johnson, W. C. (1987). Prediction of the tertiary structure of the α–subunit of tryptophan synthase. *Proteins,* **2**, 221–224.

Islam, S. A., Luo, J. C. & Sternberg, M. J. E. (1995). Identification and analysis of domains in proteins. *Prot. Eng.* **8**, 513–525.

Jin, L., Cohen, F. E. & Wells, J. A. (1994). Structure from function: screening structural models with functional data. *Proc. Nat. Acad. Sci.* **91**, 113–117.

Jones, D. & Thornton, J. (1993). Protein fold recognition. *J. Comp. Aid. Mol. Des.* **7**, 439–456.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Kabsch, W. & Sander, C. (1983). A dictionary of protein secondary structure. *Biopolymers,* **22**, 2577–2637.

Lee, J., Rieu, P., Arnaout, M. A. & Liddington, R. (1995). Crystal structure of the A domain from the α subunit of integrin CR3 (CD11b/CD18). *Cell,* **80**, 631–638.

Lemer, C., Rooman, M. J. & Wodak, S. J. (1996). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins,* **23**, 337–355.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science,* **227**, 1435–1441.

Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. & Huber, R. (1995). Crystal structure of the 20s proteasome from archaeon *T. acidophilum* at 3.4 Å resolution. *Science,* **268**, 533–539.

Lupas, A., Koster, A. J., Walz, J. & Baumeister, W. (1994). Predicted secondary structure of the 20s proteasome and model structure of the putative peptide channel. *FEBS Letters,* **354**, 45–49.

Matthews, S., Barlow, P., Boyd, J., Barton, G., Russell, R., Mills, H., Cunningham, M., Meyers, N., Burns, N., Clark, N., Kingsman, S., Kingsman, A. & Campbell, I. (1994). Structural similarity between the p17 matrix protein of HIV–1 and interferon–γ. *Nature,* **370**, 666–668.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.

Orengo, C. (1994). Classification of protein folds. *Current Opinion Structural Biol.* **4**, 429–440.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Prot. Eng.* **6**, 485–500.

Perkins, S. J., F., S. K., Williams, S. C., Haris, P. I., Chapman, D. & Sim, R. B. (1994). The secondary structure of the von willebrand factor type a domain in factor b of human complement by fourier transform infrared spectroscopy. *J. Mol. Biol.* **238**, 104–119.

Pickett, S. D., Saqi, M. A. S. & Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction. Discrimination of the β/α–barrel fold. *J. Mol. Biol.* **228**, 170–187.

Pongor, S., Hatsagi, Z., Degtyarenko, K., Fabian, P., Skerl, V., Hegyi, H., Murvai, J. & V., B. (1994). The SBASE protein domain library, release 3.0 - a collection of annotated protein-sequence segments. *Nuc. Ac. Res.* **22**, 3610–3615.

Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first–level supersecondary structure. *Proteins,* **3**, 71–84.

Rost, B. (1995). TOPITS: threading one-dimensional preditions into three-dimensional structures. *Proc. 3rd. Int. Conf. Intel. Sys. Mol. Biol.* , 314–321.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70 % accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins,* **20**, 216–226.

Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins,* **14**, 309–323.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds: An analysis of side–chain to side–chain contacts, secondary structure and accessibilty. *J. Mol. Biol.* **244**, 332–350.

Russell, R. B., Copley, R. R. & Barton, G. J. (1995). Protein fold recognition from secondary structure assignments. *Proc. 28th Hawaii. Int. Conf. Sys. Sci. IEEE Press,* **5**, 302–311.

Russell, R. B. & Sternberg, M. J. E. (1995). How good are we? *Current Biology,* **5**, 488–490.

Sheridan, R. P., Dixon, J. & Venkataraghavan, R. (1985). Generating plausible protein folds by secondary structure similarity. *Int. J. Pep. Protein Res.* **25**, 132–143.

Siddiqui, A. S. & Barton, G. J. (1995). Continuous and discontinuous domains - an algorithm for the automatic-generation of reliable protein domain definitions. *Prot. Sci.* **4**, 872–884.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance constraints. *Prot. Eng.* **6**, 605–614.

Sonnhammer, E. L. L. & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Prot. Sci.* **3**, 482–492.

Sowdhamini, R. & Blundell, T. L. (1995). An automatic method involving cluster-analysis of secondary structures for the identification of domains in proteins. *Prot. Sci.* **4**, 506–520.

Sun, S., Thomas, P. D. & Dill, K. A. (1995). A simple protein folding algorithm using a binary code and secondary structure constraints. *Prot. Eng.* **8**, 769–778.

Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Prot. Eng.* **4**, 853–870.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Current Opinion Structural Biol.* **3**, 247–259.

Zhou, M., Ravichandran, K. S., Olejniczak, E. T., Petros, A. M., Meadows, R. P., Sattler, M., Harlan, J. E., Wade, W. S., Burakoff, S. J. & Fesik, S. W. (1995). Structure and ligand recognition of the phosphotyrosine binding domain of shc. *Nature,* **378**, 584–592.