Proceedings of the 28th Annual Hawaii International Conference on System Sciences — 1995

Protein fold recognition from secondary structure assignments

R. B. Russell, R. R. Copley and G. J. Barton

Laboratory of Molecular Biophysics, The Rex Richards Building
South Parks Road, Oxford, OX1 3QU, England

Abstract

A method for finding protein folds consistent with secondary structure assignments and imposed experimental restraints is described. All possible matches between the query pattern and every member of a database of protein structural domains are generated by a comparison of secondary structure assignments. The comparison allows for errors in predicted secondary structure elements and possible variations between query and database structure. Several filters remove matches that are un-compact, that have poor β sheet bonding, that do not allow loop/turn lengths to bridge the distance between connected secondary structures, or that fail to satisfy imposed experimental restraints (e.g. disulphide bonds). The remaining matches provide a set of plausible topologies for a protein of unknown structure, which can be inspected visually or tested by experiment. A search using the src *homology 2 domain prediction finds 13 possible topologies, one being a domain from the* E. coli bio *protein known to adopt an SH2 fold. The use and development of the method are discussed.*

1 Introduction

Two of the biggest advances in protein structure prediction over the last decade have been improvements to secondary structure prediction accuracy, through the use of multiple sequence alignment and techniques such as neural networks (see references 1 & 2 for reviews) and the ability to to assess the fitness of a protein sequence to a three-dimensional (3D) structure (fold recognition; see references 3 & 4 for reviews). The former has made fairly accurate secondary structure assignments available for proteins prior to 3D structure determination, whereas the latter has suggested possible folds for more proteins than was previously possible by the comparison of sequence.

Both of these recent advances have their limitations. Not only does secondary structure prediction from multiple sequence alignment require several quite different sequences to provide an accurate prediction, but it is also only able to predict the core secondary structures for a family of proteins with confidence. Regions outside this core, which are variable in the protein sequence family, are not readily predictable. Moreover, even when the number, type and location of secondary structures is predicted correctly, variation in the ends of helices and strands is to be expected between predicted and experimental structures, or even between different experimentally determined 3D structures for proteins of the same family [5].

Protein fold recognition also has its limits. Protein 3D structure comparison (see reference [6] for a review) has found many examples of proteins adopting similar 3D folds despite no apparent sequence similarity (e.g. refs 7, 8 & 9). Comparisons have shown that many protein 3D structural similarities are slight, requiring the insertion/deletion of one or more secondary structure elements for correct alignment, and having large variations in the packing orientations of the core secondary structures. An example is shown in Figure 1. The structures of an Immunoglobulin (Ig) light chain variable domain and the N-terminal domain from haemocyanin are shown in a similar orientation. Although the sequence, function and much of the 3D structure for these two proteins is very different, they share a common core of 9 secondary structure elements (out of a possible 12). Despite having only two identical residues within this core, 51 C_α atoms can be superimposed with an RMS deviation of 2.2 Å. The lack of amino acid sequence similarity, two very large insertions (29 and 26 residues) in haemocyanin relative to the Ig domain, the difference in the lengths of many secondary structures, and differences in the residue-residue interactions that stabilise these two structures [10, 11] presents a great test of even the most robust methods.

It has been estimated that 70% of proteins of unknown 3D structure will adopt a fold similar to at least one protein of known structure [6]. However, since similar protein 3D structures often have little in common apart from their core secondary structures [11], current fold recognition methods may fail to detect many 3D structural similarities (e.g. Figure 1) prior to experimental structure determination.

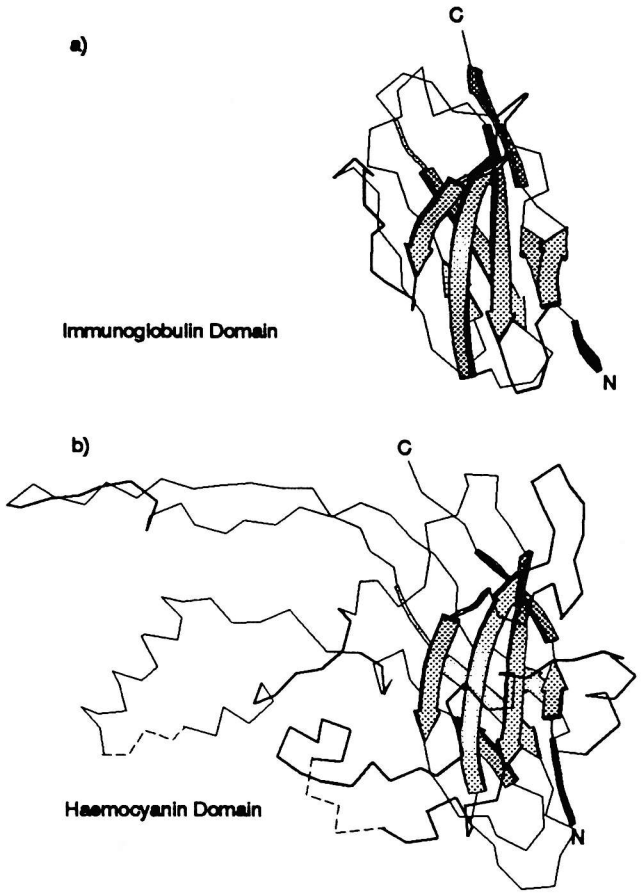

Figure 1 Example of 3D structural similarity despite no sequence or functional similarity. a) Mouse Ig light chain variable domain (PDB code 2FBJ_L, residues 1–108); b) Lobster haemocyanin (1HC1, residues 409–653) Equivalent regions [13] are shown in ribbon form; unequivalent regions as C_α trace

The recent improvements in protein secondary structure prediction provide a means to overcome the potential problems with the current methods of protein fold recognition. If the only thing necessarily common to similar protein 3D structures is the arrangement of secondary structures in space, and if accurate secondary structures are readily available (either by prediction or NMR) then a search for plausible arrangements of the secondary structures within known 3D structures, is likely to be the most successful strategy for fold recognition. To be successful, a strategy for finding such topological matches must allow for the following:

1) Variation in the ends of helices and strands, since secondary structure prediction may not predict these accurately.

2) The deletion of one or more entire secondary structure elements from the predicted secondary structure. This would allow for wrongly predicted secondary structure elements.

3) The deletion of one or more entire secondary structure elements from database structures, since alignment based structure prediction can not predict secondary structures outside the conserved core of the family of proteins correctly, and since many weak similarities of 3D structures will have the insertion of one or more secondary structures (e.g. see Figure 1; or ref. 14), or even of whole domains (e.g. ref. 15).

Due to algorithmic problems, or an extensive need for computer time, current methods of protein fold detection are likely to encounter problems dealing with requirements 2 and 3.

Sheridan *et al.* described a method for generating plausible folds from a residue–by– residue secondary structure prediction [16]. They described a simple secondary structure assignment 'mutation' matrix and used dynamic programming to find the best match between a predicted and observed string of secondary structures assignments. Spatial requirements were imposed on loops so that predicted coil regions were always able to bridge the distance required in the match. In addition, they removed any structures that were not compact or had poor β strand hydrogen bonding. Although their method was able to find possible templates for modelling a predicted α/β protein from others having no apparent sequence similarity, it is likely to be limited by the use of residue–by–residue comparisons and dynamic programing. Proteins having similar 3D structures despite no sequence similarity often have secondary structures with drastically different lengths, suggesting that residue–by–residue comparisons of predicted and experimental structures may not provide accurate alignments. Accurate alignment of distantly related protein structures may also require the deletion/insertion of entire secondary structures, which would involve large gaps in the alignment of the two sequences. The use of dynamic programming favours alignments having few gaps, and is thus likely to miss similarities such as that shown in Figure 1.

In this paper, we present a method for finding protein topologies that are consistent with predicted or NMR secondary structure assignments and with experimental information available for a protein or protein family. A query set of secondary structures and restraints is used to search a database of pro-

tein domains. All possible matches of secondary structures are generated, allowing for deletions of secondary structural elements both from the query pattern and from the database structure. These initial matches are then filtered by a series of structural criteria and experimental restraints (e.g. disulphide bonds, residues in the active site, etc.) to arrive at a collection of reasonable topologies consistent with experimental observations. The method is demonstrated by the search for topologies consistent with the prediction of the *src* homology 2 (SH2) domain [17], and is shown to be a powerful means for suggesting plausible protein folds in advance of experimental 3D structure determination.

2 Methods

A more detailed description and evaluation of the method described here will be published elsewhere. The purpose of this paper is to give a general overview of the method, and to discuss the more computationally novel ideas in some detail.

2.1 Database of protein domains

A database of unique three-dimensional structural domains was obtained from the Brookhaven Protein Databank [18]. Starting with 2048 separate chains, all sequences were compared to check for identities over their entire lengths (i.e. to remove binding studies, or duplicate structures). This left 930 non–identical chains, which were then compared using a variant of the Smith Waterman Dynamic Programming algorithm [19, 20]. 397 unique protein domains remained in the final database.

The database contains several lower quality structures (i.e. NMR, EM structures, low resolution X–ray structures, structures having only C_α atoms, etc.). During a search for plausible folds, it is advantageous to consider all structures, regardless of quality, since there is often only one, low quality example of a unique protein fold (e.g. phaseolin, PDB code 1PHS). However, when deriving parameters (see below) it is better to use higher quality structures. For this reason, a sub–database of 256 higher quality domains was created, containing only those structures determined by X–ray crystallography refined and of a resolution of 2.5 Å or better.

2.2 Predicted secondary structures

Multiple sequence alignments can provide much more than a prediction of secondary structures. Gaps within the alignment correspond quite accurately to loops in the core structure common to the family of proteins. Conservation of hydrophobicity or polar properties can indicate those residues buried within the core of the protein or exposed to solvent. Conservation of other residues, taken with knowledge of the known function of the protein (e.g. from SDM or other experiments) can suggest residues likely to be near to each other in the native fold (e.g. within the active site, or a binding site). In addition to predicted secondary structures, any method to predict topology ought to make use of such additional information to restrict possible folds [21, 22, 23].

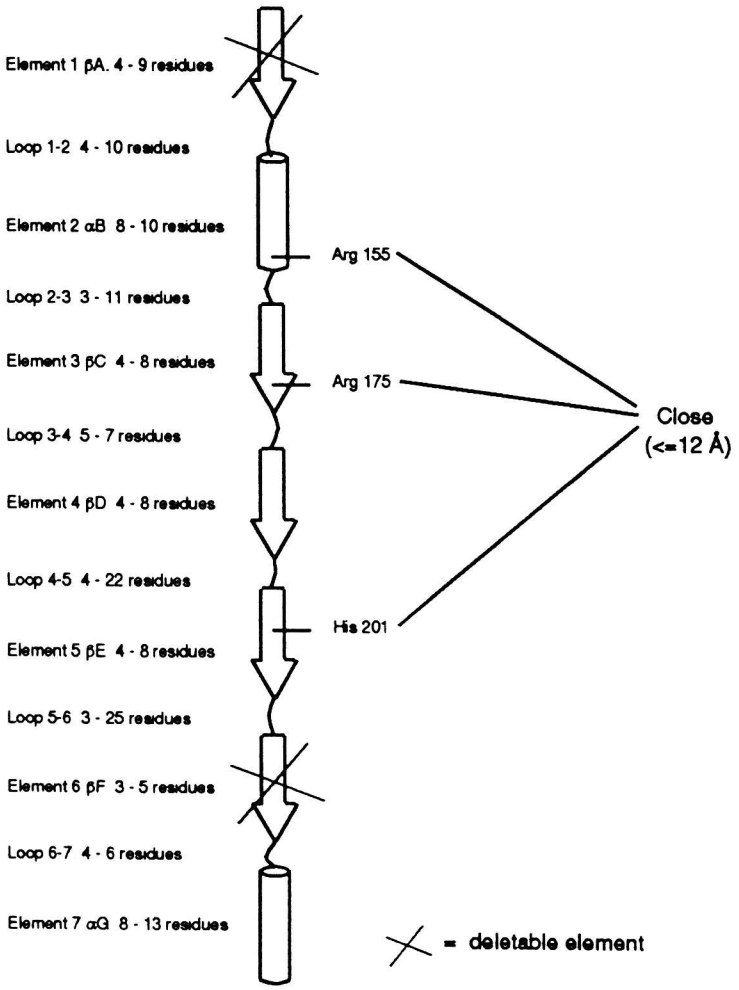

Figure 2 Predicted structure for the SH2 domain as described by Russell *et al.* [17]

Figure 2 shows an example of the information contained within a secondary structure prediction pattern

from our prediction of the *src* homology 2 (SH2) domain [17]. In addition to predicted secondary structures (type, minimum and maximum length), there are restraints as to the minimum and maximum loop lengths between secondary structures. From experimental observations, and residue conservation, three residues Arg 155, Arg 175 and His 201 appeared to be involved in phosphotyrosine binding [24, 17] prior to 3D structure determination. It is sensible to introduce a weak distance restraint by requiring that the coordinates corresponding to these positions be near to each other in any predicted topology.

2.3 Experimental secondary structures

Secondary structures were calculated using the method of Kabsch & Sander (DSSP) [25], and all 397 domains were split into separate helices and strands. DSSP α and 3_{10} helix were considered to form a helix (H) when a continuous run of more than four residues had either of these conformations. DSSP β sheet and β bridge were considered to form a strand (B) when more than two residues had either of these conformations. The remaining secondary structure classifications were defined as coil (-). For 3D structures containing only C_α coordinates, the method of Richards & Kundrot (DEFINE) [26] was used to define helix and strand. Residues assigned as β sheets by DEFINE were defined as strand, DEFINE α or 3_{10} helix were defined as helix, and the remaining classifications as coil. The length requirements for strand and helix were as for DSSP assignments. 3D coordinates for strands and helices were used to derive a set of axial coordinates as described by Richards & Kundrot [26]. Ideal curves fitted through each secondary structure element are used to provide a single axial coordinate for each amino acid. To allow for variation in the ends of secondary structure elements, axial coordinates were calculated for 20 additional residue positions in the direction of their N- and C- terminii. This strategy provides axial coordinates in situations where a query secondary structure element is longer than an observed element, or where query elements hang over the end of database elements (see below).

2.4 Matching secondary structure assignments

The first stage of the mapping procedure involves searching for a match of a predicted (or *query*) set of secondary structure elements with known (or *database*) patterns of secondary structure elements. The simplest procedure would be to look for all exact matches (i.e. matches having the exact set of secondary structures in exactly the right order) of a query in a database structure. However, as said above, alignment of protein 3D structures often requires the insertion/deletion of one or more secondary structural elements. The similar 3D structures shown in Figure 1 provide an example. By representing each secondary structure domain as either H (helix) or B (strand), the structural alignment shown in Figure 1 can be written:

```
Haemocyanin BBBHBBBBB--BBBHB-BHBB
Ig domain   BB----BBBBB-B---BB-BB
```

Despite the possibility for many other possible matches of the Ig and haemocyanin secondary structures, the only structurally viable match is that shown above, requiring the deletion of three secondary structures from the Ig domain and nine from the Haemocyanin domain. Any method to search for possible topologies for an arrangement of secondary structures should allow for the possibility of such insertions/deletions.

A matching algorithm was developed to generate all possible alignments of query with experimental secondary structural elements allowing for insertions/deletions. A user supplied maximum number of deletions from the query and database structures is used with a recursive routine to find all matches consistent with these parameters. Figure 3 shows a skeleton code representation of the algorithm.

The first element in the query is passed along the string of elements in the database structure until a match is found. After a match has been found, the routine is called again (i.e. it calls itself) with the query and database elements one after (i.e. C-terminal to) the position of the previous match. The process is repeated until either query or database sequences have reached their end. Whenever a minimum number of matches has been found, the match is saved. Deletions in the database sequence are allowed implicitly by allowing the query element to match with any database element after the last match. Deletions are allowed in the query by attempting to delete every position at the start of each call of the recursive routine

(see Figure 3).

```
string Q,D                          query and database strings (eg. HBBHBBB)
binary array Qmatch, Dmatch         list of matches (none initially)
binary  Qbits, Dbits                current set of bits matched (all 0 initially)
integer counter,                    number of matches (0 initially)
        Qpos, Dpos,                 current positions for query and database (both 1 initially)
        total_Qdels,total_Ddels,    current number of deletions to query and database both 0 initially)
        total_matched,              current number of matched bits
        max_Qdels,max_Ddels,        maximum number of deletions to query /database (user defined)
        min_match                   minimum number of matches required to save match (user defined)
SECONDS is array of binary numbers from 1 to N where SECONDS(i) corresponds to the binary number
        which as the ith bit turned on (eg. SECONDS(3) = 0010 0000 0000 0000 ... )

main program
        Set: Qpos = 1, Dpos = 1, total_Qdels = 0, total_Ddels = 0,
             total_matched = 0, Qbits = 0, Dbits = 0, counter = 0
        User defined: max_Qdels, max_Ddels,min_match
        Call routine: comparing(Q,Qpos,Qbits,total_Qdels,D,Dpos,Dbits,total_Ddels,total_matched)
        Filter matches as described in text

recursive procedure comparing(Q,Qpos,Qbits,total_Qdels,D,Dpos,Dbits,total_Ddels,total_matched) {
    local integer i, old_total_Ddels
    Qlen = length of string Q  Dlen = length of string D
    If NOT redundant hit AND total_Qdels < max_Qdels then { try and delete the positions
        If total_matched >= min_match then {          if we have enough matches, save the result
          Qmatch(counter) = Qbits  Dmatch(counter) = Dbits
          counter = counter + 1
        } else {
          total_Qdels = total_Qdels + 1                    delete the position
          Qpos = Qpos + 1
          comparing(Q(Qpos),Qpos,Qbits,total_Qdels,D,Dpos,Dbits,total_Ddels)  call routine again
          total_Qdels = total_Qdels - 1     Qpos = Qpos - 1    un-delete the position
        }
      }
      old_total_Ddels = total_Ddels                    save old database deletion value
      for i = 1 to (Dlen - Qlen + max_Qdels -total_Qdels + 1) do {
        If Q(0) = D(i) AND total_Ddels <= max_Ddels then {
        total_matched = total_matched + 1
          Qbits = Qbits bitwise OR SECONDS(Qpos)     turn matched bits on
          Dbits = Dbits bitwise OR SECONDS(Dpos+i)
          Qpos = Qpos + 1  Dpos = Dpos + i           update positional pointers
          If NOT redundant hit AND total_matched >= min_match then {
            Qmatch(counter) = Qbits  Dmatch(counter) = Dbits  counter = counter + 1
          }                                          save the result
          If  Qlen > 1 AND Dlen > 1 then {
           comparing(Q(2),Qpos,Qbits,total_Qdels,D(i),Dpos,Dbits,total_Ddels)  call routine again
        }
          Dpos = Dpos - i  Qpos = Qpos -1            return pointers to previous values
          Dbits = Dbits bitwise XOR SECONDS(Dpos+i)   turn matched bits off again
          Qbits = Qbits bitwise XOR SECONDS(Qpos)
          total_matched = total_matched - 1
        }
        If total_matched > 0 then  total_Ddels = total_Dels + 1  if no match then count as deletion
      }
      total_Ddels = old_total_Ddels                  reset old database deletion value
      return to last recursive level or end          got to end of string D  go back to last level
}
```

Figure 3 Skeleton code for the matching algorithm described in the text

For speed and efficiency, matches between query and database are stored in a binary format. Two binary numbers describe each match: one for the query and one for the database structure. Each secondary structure element, in both query and database patterns, is represented by a single bit: 1 indicates that the element is considered in the match, 0 that it is not. For each match, the pair of binary numbers will have the same number of 1s, and an alignment of secondary structure elements can be obtained by matching the ith included query secondary structure (i.e. bit set to 1) with the ith included database secondary structure. For example, if only one deletion in the query structure is allowed, a query secondary structure pattern HHBBH would have 6 matches with a database secondary structure pattern HBBBH, which are represented by the following pairs of binary numbers:

```
     HHBBH HBBBH
1  01111 11101
2  01111 11011
3  01111 10111
4  10111 11101
5  10111 11011
6  10111 10111
```

The matches are shown in the order in which they are found by the algorithm. The binary representation means that many matches of query to database can be easily stored for further analysis (i.e. Filtering, see below), and that simple bitwise tests (AND, OR, exclusive OR, etc.) can be used for fast sorting and filtering.

The method described here has advantages over alternative matching methods, such as dynamic programming or stochastic algorithms, since it easily copes with subtle variations (or sub–optimal alignments). Since all matches of secondary structure correspond directly to a 3D structure, it is important to consider such subtle variations. For example, two possible alignments between query and database might be represented as:

```
            Alignment 1    Alignment 2
Query       BBHBB--BH--    BBHBBB--H--
Database    -BHBBBHBHBB    -BHBBBHBHBB
```

The difference between the two alignments when displayed in a one–dimensional form is marginal. However, a consideration of the two alternative 3D structures might show one to have an isolated β strand (i.e. not properly hydrogen bonded), making it a nonsensical match. If such possibilities are to be considered further, then the method must allow for such subtle variations.

The matching method has some computational restrictions, since some comparisons can lead to a huge number of possibilities (e.g. 7 predicted strands matching to 20 observed strands). Provided that the number of deletions in the query structure is kept to a minimum, then the number of matches which need to be considered can be easily kept below 1 000 000. Since the object of this method is to detect plausible arrangements of a query set of secondary structures in a database of known 3D structure, it is not practical to allow large numbers of deletions in the query

(say more than a third of the elements in the pattern). Deleting too much of the query pattern defeats the object of a search. On the other hand, it is advantageous to allow for numerous deletions from the database structure, since this allows for large insertions (even those containing several secondary structures) to be tolerated during the alignment of query and database structures. The number of initial matches is also kept to a minimum by the use of a database of domains rather than whole 3D structures (i.e. the average database domain size is kept to a minimum).

2.5 Filters

The routine described in the previous section will return all possible matches of a query to a database allowing for a particular number of deletions. However, many of these will be nonsensical, since they will consist of uncompact structures (e.g. separate parts of a larger structure), or will have secondary structures too far away from any others to form a sensible fold. In addition, some matches may not agree with experimental information for the protein. For example, the orientation of secondary structures may prevent the formation of disulphide bridges, or of active/binding sites. Nonsensical matches and those not satisfying any imposed experimental restraints are removed by a series of filters discussed below.

2.6 Compactness

To remove grossly uncompact structures, a coarse filter is applied based on radius of gyration (R_g).

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} m_i(|\vec{r_i} - \vec{R_o}|)^2}{\sum_{i=1}^{n} mi}}$$

where n is the number of secondary structures in the match, m_i is the molecular weight of the ith secondary structure, $\vec{r_i}$ is the vector describing he position of the centre of mass for the ith secondary structure, and $\vec{R_o}$ is the centre of mass for the entire match.

An analysis of high quality domains shows that R_g obeys the equation:

$$R_g \leq 2.36L^{0.366} + 1.00$$

where L is the total length of the domain (in residues). Any matches having an R_g greater than this ideal value are ignored.

Since matches contain only secondary structural elements and lack any loops, an approximation of the loops is obtained by considering those loops in the database structure immediately C–terminal to each matched secondary structural element.

2.7 Variation in secondary structure lengths and positions

The variation in the secondary structures of a family of homologous proteins suggests that even the best secondary structure predictions fail to predict the precise starts/ends of secondary structures within a protein family [5]. In addition, proteins having similar 3D structures in the absence of any sequence similarity may have secondary structures of very different lengths. Thus even if query and database experimental structures are matched correctly, the precise positions of individual residues in the query sequence (i.e., when mapped on to the database structure) can not be known. It is therefore necessary to allow for a range of possible positions of a query residue within a database secondary structure when imposing a filter based on distance restraints.

A range of possible database positions are defined for any sequence position on on the query. Given a residue position x on a predicted secondary structure, and a minimum and maximum length (in residues) for that secondary structure (L_{min} and L_{max}, a range of values or positions of x can be determined by using the matched database secondary structure length (L_{obs}),

$$x_{min} = \text{minimum of } (L_{obs} - L_{max}, 0 - h) + x$$
$$x_{max} = \text{maximum of } (L_{obs} - L_{min} + h, 0) + x$$

where h is an 'overhang' value, or the number of residues in the query secondary structure allowed to hang off either the N– or C– terminii of the database secondary structure (set to 2). Any query position x can have database positions between x_{min} and x_{max} (between -20 and $L_{obs} + 20$), and these positions correspond directly to axial positions within the database 3D structure. For example, if $h = 2$, $x = 4$, $L_{obs} = 8$, $L_{min} = 7$, and $L_{max} = 10$, then the the fourth position in the predicted secondary structure can be anywhere within positions 2 and 6 of the observed secondary structure. Large differences between predicted and observed lengths lead to greater flexibility in the possible positions. For example, if $h = 2$, $x = 4$, $L_{obs} = 15$, $L_{min} = 5$, and $L_{max} = 7$, then the range of values is between 2 and 14. To allow maximum leniency, the minimum distance between two positions (x_1 and x_2), found by considering all combinations of both ranges, is used during filters using distance restraints (see Loop and Distance filtering below).

2.8 Loops between adjacent secondary structures

Secondary structure predictions, whether for a single sequence or a family, provide information as to the location of loops within the tertiary structure. Allowing for variation in the ends of secondary structures, and for different sequence lengths within loop or turn regions among a family of protein sequences means that one can assign a minimum and a maximum coil length to link the end of one predicted secondary structure to the start of another (the *end–to–start* distance). Analysis of the database of high quality domains shows that the maximum inter–molecular protein distance, D, that can be crossed by a number of residues, N is approximately:

$$D = 2.5N + 4.0 \text{ Å}$$

For example, 3 residues can only bridge a distance of $\approx$ 11.5 Å, whereas 10 residues can bridge distances up to $\approx$ 29 Å. The minimum number of residues thought to connect any two predicted secondary structures thus places restrictions on the topologies allowed in the form of distance restraints. In order for a match of predicted and database secondary structures to be a plausible fold, adjacent predicted secondary structures matched to database structures must have enough residues to cross the minimum end–to–start distance. Any matches with adjacent secondary structures in orientations with end–to–start distances that are too long for the minimum predicted loop length are ignored.

2.9 β sheets with poor hydrogen bonds

When a match of predicted and database secondary structure includes β strands, it is necessary to ensure that no β strands lack hydrogen bonding partners. An analysis of C_α - C_α contacts within strands in the high quality domains shows that all β strands of length L have at least $L + 2$ C_α - C_α contacts less than 6 Å with other β strands. Matches having any β strands with fewer than $L + 2$ such contacts are ignored.

2.10 Essential secondary structures

Apart from a prediction of secondary structure elements, sequence analyses and other biochemical studies provide further restraints on the possible folds for a particular structure. For example, SDM experiments may provide the location of active site or other residues important to the function of the protein. If query to database matches are found that lack the secondary element containing such a residue, then they may be ignored. In addition, many secondary structure prediction methods assign different levels of confidence to different regions of a prediction [17, 27, 28]. If particular secondary structure elements are strongly predicted, then one may wish to remove matches lacking such elements.

A filter can be applied to remove all those matches lacking essential secondary structure elements. For example, the pattern shown in Figure 2 has two 'deletable' secondary structures (elements 1 and 6); the rest (2,3,4,5 and 7) are deemed essential. Elements 3,4,5 and 7 were strongly predicted, and elements 2,3 and 5 contain residues that were thought to be involved in phosphotyrosine binding.

2.11 Distance restraints

Biochemical studies done prior to experimental 3D structure determination can also suggest residues likely to be close together in space. SDM may suggest residues that act together in the active/binding site. For proteins containing disulphide bonds, the disulphide bonding pattern may be known, since this is often determined during protein sequencing experiments. Such distance restraints can be used as another filter; any matches between query and database with matched secondary structures not capable of satisfying such restraints are ignored.

Using the minimum and maximum positions values for two residues x_1 and x_2 as describe above, matches can be filtered by only accepting those having axial positions within a specified cutoff. Analysis of several disulphide containing proteins of known 3D structure shows that axial positions for the cysteine residues involved are within $\approx$ 9.5 Å (results not shown). Active site residues are less constraining, typically having axial coordinates within $\approx$ 12 Å (results not shown). Thus, for the pattern shown in Figure 2, the residues thought to be involved in phosphate binding are restricted to have axial positions within 12 Å.

2.12 From mapping to modelling

If a query secondary structure prediction pattern produces a set of possible alternative topologies, then the next step is to select the best possible template for homology modelling. For small numbers of topologies, this can be done quite easily by eye, since the restraints used to filter the matches can leave impossible structures (e.g. requiring loops through

secondary structures, having active site residues in implausible orientations, etc.). Once obviously bad matches have been removed, an alignment of query to database may be obtained by a restrained threading algorithm, which gets the best match between sequence and structure, but does not violate any of the restraints discussed above, and ensures correct matching of query and database secondary structures. A method to obtain such alignments is being developed.

2.13 Program details

The program to perform the above matching and filtering was written in C and developed on a Sun SPARCstation ELC. The program has not yet been optomised for speed. Loading the database of 397 structures takes 15 minutes of CPU time, and comparing the pattern shown in Figure 2 to the database, allowing for 2 deletions from the query and an unlimited number of deletions from the database structures, takes approximately 5 minutes of CPU time on a Sun SPARCstation ELC.

3 Results and Discussion

3.1 Search with the SH2 domain

To give a preliminary demonstration of the method, the prediction pattern shown in Figure 1 was compared to the database of 397 domains, and passed through the six filters discussed above. Figure 4 shows how the number of matches dropped during the various filtering stages. Allowing for 2 deletions in the predicted secondary structure and an unlimited number of deletions in any database structure, comparison of secondary structure assignments gave 504 406 initial matches. The coarse radius of gyration filter and loop filtering bring this number down by a factor of ten to 68 410 compact matches. Removing those matches with poor β sheet bonding leaves only 8 622. Requiring elements 2,3,4,5 and 7 to be in any match leaves only 768, of which only 50 satisfy the three distance restraints. Removing those remaining matches contained entirely within another (i.e. redundant) left 37 matches, and only 34 of these were found to be unique when investigated using a protein structure comparison algorithm [13] (i.e. repeated folds were removed).

The 34 final structures were investigated using molecular graphics, and the TOPS program [29]. Of the remaining 34, 21 had bad β sheet topologies, where adjacent predicted strands were parallel (i.e. loops were required to cross the entire domain). Removing these left only 13 possible topologies, of which 7 contained a single β sheet and 6 contained two β sheets. These final matches are shown as topology diagrams in Figure 5. Among the 13 matches remaining is the portion of the structure of the *E. coli* biotin operon protein (BirA) that is known to be similar to the SH2 domain [14]. Also in the list are other possible topologies consistent with the restraints given, such as portions of malate and lactate dehydrogenase, or glutathione reductase. If such alternative topologies are available for proteins of unknown structure, they can be investigated using molecular graphics, or tested by experiment, such as site–directed mutagenesis, or even used for molecular replacement models during molecular replacement.

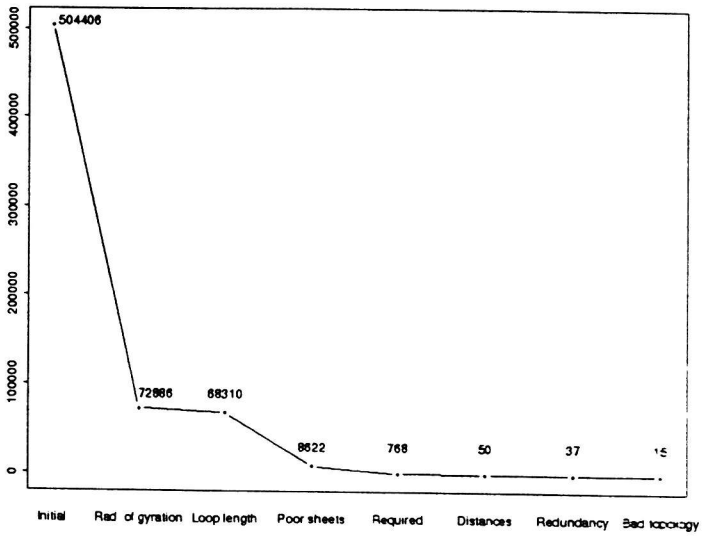

Figure 4 How the number of matches decreases as the various filters are applied during a search with the SH2 domain pattern (Figure 2)

3.2 Using the method

In practice the method described here is used iteratively. One must try searches with a particular set of restraints, and then relax or tighten these restraints depending on the matches obtained. If one is uncertain as to whether particular elements are helices or strands, then multiple searches can be performed, and the results compared. In this way, the matches found can be used to modify the secondary structure assignment.

3.3 Future developments

A method for automating the filtering done above by eye based on bad β sheet topology is under development. By analysis of known protein 3D structural domains, a ranking system is being developed based on how well members in a list of topologies meet general topological requirements. For example, structures having many secondary structures adjacent on the sequence and parallel in the 3D structure can be considered poor, since this is rare among known 3D structural domains.

A method for obtaining the best alignment of predicted sequence to mapped structure is also under development. By calculating the accessibility of residues within each match, the best alignment can be obtained by aligning predicted amphipathicty (i.e. conserved hydrophobic, conserved charge, etc.) with observed residue exposure/burial and by obeying any other restraints imposed on the match (i.e. disulphides, or other distance restraints).

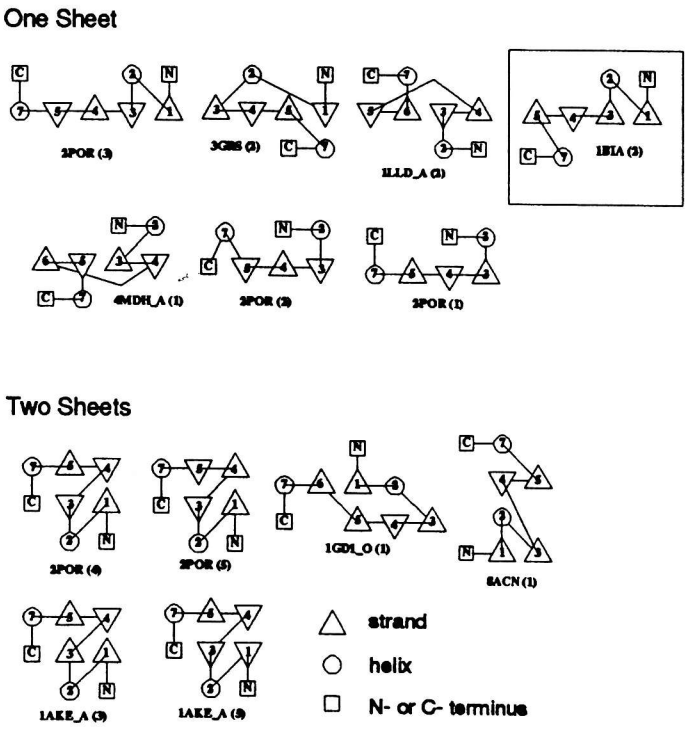

Figure 5 TOPS [29] diagrams of the final 13 hits obtained during the search with the SH2 domain pattern. Numbers inside secondary structures are as for the predicted structures defined as Figure 2.

4 Conclusions

The method presented here is a potentially powerful tool for protein structure prediction. Given a set of secondary structures, the method quickly provides a set of plausible folds that are both sensible structures, and consistent with imposed experimental restraints. The method promises to be an effective link between secondary structure prediction, fold recognition and homology modelling.

Acknowledgements

We thank Professor L.N. Johnson for providing a stimulating working environment. RBR and GJB thank the Royal Commission for the Exhibition of 1851 and the Royal Society for support. RRC is funded by a Medical Research Council (UK) studentship and is member of St. Catherine's College, Oxford.

References

[1] B. Rost, R. Schneider, and C. Sander. *Trends Biochem. Sci.*, 18:120–123, 1993.

[2] S. A. Benner and D. L. Gerloff. *FEBS Lett.*, 325:29–33, 1993.

[3] S. J. Wodak and M. J. Rooman. *Curr. Opin. Struct. Biol.*, 3:247–259, 1993.

[4] J. U. Bowie and D. Eisenberg. *Curr. Opin. Struct. Biol.*, 3:437–444, 1993.

[5] R. B. Russell and G. J. Barton. *J. Mol. Biol.*, 234:951–957, 1993.

[6] C. Orengo. *Curr. Opin. Struct. Biol.*, 4:429–440, 1994.

[7] L. Holm and C. Sander. *Nature*, 361:309, 1993.

[8] M. B. Swindells, C. A. Orengo, D. T. Jones, L. H. Pearl, and J. M. Thornton. *Nature*, 362:299, 1993.

[9] R. B. Russell. and G. J. Barton. *Nature*, 364:765, 1993.

[10] B. Hazes and W. G. J. Hol. *Proteins: Struct. Funct. Genet.*, 12:278–298, 1992.

[11] R. B. Russell and G. J. Barton. *J. Mol. Biol., In Press.*, 1994.

[12] C. A. Orengo, T. P. Flores, W. R. Taylor, and J. M. Thornton. *Protein Engng.*, 6:485–500, 1993.

[13] R. B. Russell and G. J. Barton. *Proteins: Struct. Funct. Genet.*, 14:309–323, 1992.

[14] R. B. Russell and G. J. Barton. *Nature*, 364:765, 1993.

[15] J. L. Martin, J. C. A. Bardwell, and J. Kuriyan. *Nature*, 365:464–468, 1993.

[16] R. P. Sheridan, J.S Dixon, and R. Venkataraghavan. *Int. J. Peptide Protein Res.*, 25:132–143, 1985.

[17] R. B. Russell, J. Breed, and G. J. Barton. *FEBS Lett.*, 304:15–20, 1992.

[18] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanovichi, and M. Tasumi. *J. Mol. Biol.*, 112:535–542, 1977.

[19] T. F. Smith and M. S. Waterman. *J. Mol. Biol.*, 147:195–197, 1981.

[20] G. J. Barton. *CABIOS*, 9:729–734, 1993.

[21] F. E. Cohen and M. J. E. Sternberg. *J. Mol. Biol.*, 137:9–22, 1980.

[22] S. A. Benner and D. Gerloff. *Adv. Enz. Reg.*, 31:121–181, 1990.

[23] D. A. Clark, G. J. Barton, and C. J. Rawlings. *J. Mol. Graph.*, 8:94–107, 1990.

[24] C. A. Koch, D. Anderson, M. F. Moran, C. Ellis, and T. Pawson. *Science*, 252:668–673, 1991.

[25] W. Kabsch and C. Sander. *Biopolymers*, 22:2577–2637, 1983.

[26] F. M. Richards and C. E. Kundrot. *Proteins: Struct. Funct. Genet.*, 3:71–84, 1988.

[27] B. Rost and C. Sander. *J. Mol. Biol.*, 232:584–599, 1993.

[28] C. D. Livingstone and G. J. Barton. *Int. J. Peptide Protein Res.*, 44:239, 1994.

[29] T. P. Flores, D. S. Moss, and J. M. Thornton. *Proten Engng.*, 7:31–38, 1994.