Prediction of Protein Secondary Structure and Active Sites using the Alignment of Homologous Sequences

The prediction of protein secondary structure (α-helices, β-sheets and coil) is improved by 9% to 66% using the information available from a family of homologous sequences. The approach is based both on averaging the Garnier *et al.* (1978) secondary structure propensities for aligned residues and on the observation that insertions and high sequence variability tend to occur in loop regions between secondary structures. Accordingly, an algorithm first aligns a family of sequences and a value for the extent of sequence conservation at each position is obtained. This value modifies a Garnier *et al.* prediction on the averaged sequence to yield the improved prediction. In addition, from the sequence conservation and the predicted secondary structure, many active site regions of enzymes can be located (26 out of 43) with limited over-prediction (8 extra). The entire algorithm is fully automatic and is applicable to all structural classes of globular proteins.

More than 3700 protein sequences are known (e.g. see Barker *et al.*, 1986) and this wealth of biological information has highlighted the need for accurate and automatic methods to predict protein conformation and function from primary structure (for reviews, see Sternberg, 1986; Taylor, 1986*a*). However, recently Kabsch & Sander (1983) have evaluated the accuracies of three widely used and general methods of predicting secondary structure. They considered more structures than the data set on which the algorithms were developed and found that the accuracies for a three-state prediction (α-helix, β-sheet and coil) were: 56%, Robson and co-workers (Garnier *et al.*, 1978); 50%, Chou & Fasman (1978); and 59%, Lim (1974).

Since these three methods were developed in the 1970s, there have been several new algorithms reported. Taylor & Thornton (1983, 1984) started with Robson's approach as it is both probabilistic and simple to program (see below) and improved it by an average of 7·5% when applied to the α/β class of proteins. A length-dependent template for the β-strand/α-helix/β-strand motif modified the likelihood of α and β structure. Another recent approach by Cohen *et al.* (1983, 1986) is based on pattern recognition of specific residue types and has been applied to the α/β class of proteins and to predict turns in all structural classes. However, the work of both Taylor & Thornton (1983, 1984) and Cohen *et al.* (1983, 1986) requires the assignment of structural class (Levitt & Chothia, 1976) and this still cannot be reliably obtained.

With the increase in number of known protein structures (Bernstein *et al.*, 1977) several groups (Sweet, 1986; Nishikawa & Ooi, 1986; Levin *et al.*, 1986) have recently explored a different approach. These methods are based on recognizing a sequence relationship between a segment of the polypeptide chain of the unknown structure with a sequence and conformation data base from the known structures. The accuracies for a three-state prediction range between 59% and 63%.

Today when one wishes to predict the secondary structure of a protein, one often has sequences from a family of homologous molecules. This provides additional information that needs to be incorporated into predictions. In this paper an algorithm is presented that uses the observation that when sequences are aligned the regions of insertions and low sequence conservation tend to occur in the loop regions connecting the regular secondary structures. This aspect is quantified and modifies the standard algorithm of Garnier *et al.* (1978) to yield an improved prediction. In addition, residues central to the activity of an enzyme are conserved in the family of homologous molecules. From the aligned sequences and predicted secondary structure, an algorithm is developed to identify potential functionally important residues (FIRs)†.

The first step of the algorithm is to obtain an alignment of all the sequences. A standard method to align two protein (or nucleic acid) sequences is the dynamic programming approach of Needleman & Wunsch (1970). A matrix of similarity (identity, chemical property or observed substitution) between pairs of amino acids is chosen and the algorithm establishes an alignment of the two sequences including insertions that yields the best score. A previous study (Barton & Sternberg, 1986) has evaluated the accuracy of sequence alignments on the basis of a benchmark obtained from structural superpositions of the two molecules. It was shown that if the score is greater than 6 standard deviations from the mean score for random sequences, then the residues in secondary structures generally are aligned to >75% accuracy. Thus features in the aligned sequences

† Abbreviation used: FIR, functionally important residues.

0022-2836/87/120957-05 $03.00/0

could be used to locate secondary structures. Although this algorithm could be generalized to align many sequences simultaneously, computer memory and time requirements are prohibitive (e.g. see Murata *et al.*, 1985). Therefore, a simpler approach for multiple sequence alignment was developed.

The method applies the Needleman & Wunsch (1970) algorithm at each stage. Firstly, sequence 2 is aligned with sequence 1, then sequence 3 is aligned with the alignment of 1 and 2 by using a similarity score obtained from the mean score at each position for 3 *versus* 2 and 3 *versus* 1. This procedure is then continued for sequences 4 to N. Once all N sequences have been aligned, sequence 1 is realigned to the 2 to N sequences; then 2 against 1, 3, 4, . . ., N, etc. One further pass is required to produce an alignment that appears consistent by visual inspection. The algorithm, without any optimization, requires 90 minutes of VAX 11/750 c.p.u. time to align 60 sequences of about 110 residues. This alignment procedure was applied to 11 families of sequences (Table 1) obtained from the Protein Information Resource data bank (Barker *et al.*, 1986), which were chosen so that one member of the family had a crystallographic secondary structure assignment (Bernstein *et al.*, 1977).

To obtain a benchmark against which improvements can be assessed, a standard Robson prediction (Garnier *et al.*, 1978) was performed on the protein with known secondary structure. In outline, the Robson method evaluates the likelihood of residue i being in an α-helix by the addition of the empirical α-helix-forming propensities of 17 residues from $i-8$ to $i+8$. Similarly, the likelihoods of the residue being in a β-sheet, turn or coil are evaluated and the conformation state with the maximum likelihood is predicted for residue i.

One approach for incorporating the information from the family of sequences, suggested when the Robson algorithm was developed, is to average the α, β, coil and turn parameters for all the aligned residues at each position. In our algorithm each insertion is scored as 0·0, which is a neutral value as Robson propensities are both positive and negative numbers.

There is, however, more information about secondary structure available from aligned sequences than that simply obtained by averaging the residue propensities. The crystal structures of protein families show that sequence insertion and sections of high sequence variability occur in loop regions between secondary structures (e.g. see Greer, 1981). In addition, residues involved in secondary structure packing tend to be hydrophobic (Lesk & Chothia, 1980). We have quantified the extent of sequence conservation at a position i along the chain by a "conservation number" C_i, which ranges from 0 to 1. C_i is based on a representation of the Venn diagram of the chemical properties of the amino acids (Taylor, 1986*a*,*b*). The aim is to have a high conservation number when similar chemical types of amino acids occur at a position, and a low value when there is high variability of residues or an insertion. Thus a conserved residue scored 1·0 and substitutions between residues with the same chemical properties 0·9. For chemically different amino acids 0·1 was subtracted from 0·9 every time one of the chemical properties was different.

Table 2 lists the amino acids and assigns a yes or no value to ten chemical properties. Gaps and unknown residues are modelled by assigning a yes value for each property. If an invariant residue occurs at a position then C_i is 1·0. When a set of residues occur at a position, each property is considered in turn and if one amino acid differs in

Table 1
Prediction of secondary structure

					Accuracy of prediction (%)		
Reference protein		No. of residues	No. seq. aligned	Average conservation, C_{av}	Robson on reference	Robson on multiple	Robson + conservation
(1) Haemoglobin (β-chain)	(α/α)	146	63	0·4	66·0	64·2	71·2
(2) Cytochrome *c*	(α/α)	103	57	0·4	52·4	57·2	59·2
(3) Myoglobin	(α/α)	153	28	0·5	67·0	68·6	71·8
(4) Immunoglobulin Fab V_H	(β/β)	117	59	0·2	60·6	67·5	68·3
(5) Kallikrein	(β/β)	223	5	0·5	54·3	60·3	63·3
(6) Lactate dehydrogenase	(α/β)	329	7	0·6	56·0	58·9	65·3
(7) Dihydrofolate reductase	(α/β)	162	11	0·2	53·0	52·0	55·5
(8) Triose phosphate isomerase	(α/β)	247	5	0·7	61·1	66·3	69·6
(9) Phospholipase A_2	($\alpha+\beta$)	123	26	0·4	44·5	53·4	61·8
(10) Ribonuclease S (bovine)	($\alpha+\beta$)	125	22	0·6	64·5	66·9	73·4
(11) Lysozyme (human)	($\alpha+\beta$)	130	6	0·8	53·1	59·2	68·4
Average		—	—	—	57·5	61·3	66·1

The Brookhaven files (Bernstein *et al.*, 1977) corresponding to the protein numbers in the Table are: (1) 2MHP; (2) 3CYT; (3) 1MBN; (4) 3FAB; (5) 2PTN; (6) 4LDH; (7) 1DFR; (8) 1TIM; (9) 1BP2; (10) 1RNS; (11) 2LYZ. The structural class of the protein is indicated. The value of C_{av} and the number of aligned sequences provide some guide as to the extent of sequence variation, and inspection of the results shows that there is no direct relationship between these values and the improvement in secondary structure prediction. However, the proteins were chosen so that there was enough variation in sequence to provide additional information, but the sequences were homologous enough that they should have very similar secondary structures.

Table 2
Properties of amino acid residues

Properties: Amino acid	Hydrophobic	Positive	Negative	Polar	Charged	Small	Tiny	Aliphatic	Aromatic	Proline
Ile	Y							Y		
Leu	Y							Y		
Val	Y					Y		Y		
Cys	Y					Y				
Ala	Y					Y	Y			
Gly	Y					Y	Y			
Met	Y									
Phe	Y								Y	
Tyr	Y			Y					Y	
Trp	Y			Y					Y	
His	Y	Y		Y	Y				Y	
Lys	Y	Y		Y	Y					
Arg		Y		Y	Y					
Glu			Y	Y	Y					
Gln				Y						
Asp			Y	Y	Y	Y				
Asn				Y		Y				
Ser				Y		Y	Y			
Thr	Y			Y		Y				
Pro						Y				Y
Asx				Y						
Glx				Y						
Gap	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Unknown	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Y, yes value.

the yes or no assignment of chemical property, a count (P) is incremented. The conservation number is calculated as:

$$C_i = 0{\cdot}9 - 0{\cdot}1 \times P.$$

If $P = 10$, then C_i is set to 0·0 rather than $-0{\cdot}1$. Thus if Ile and Leu only occur at a position, then $P = 0$ and $C_i = 0{\cdot}9$ reflected the chemical similarity of the residues. If Ile and Glu occur, $P = 5$ and $C_i = 0{\cdot}4$. The effect of a gap is to lower the value of C_i. Thus Ile and a gap yield a value for P of 7 and $C_i = 0{\cdot}2$. If Ile, Glu and a gap occur than $P = 10$ and $C_i = 0{\cdot}0$.

After C_i is calculated for each position, C_i is averaged over three residues ($i-1$, i, $i+1$), to yield a "smoothed conservation number" CS_i. When CS_i is plotted along the sequence (Fig. 1) and the regular secondary structures marked on the plot, the lower values of CS_i occur most frequently in the loop regions. Thus this plot alone is helpful in suggesting the locations of secondary structures.

These observations can be quantified to improve secondary structure prediction. The aim is to penalize the prediction of secondary structure where there is a low conservation number. First, the average conservation number over the entire sequence (C_{av}) is obtained. The smoothed conservation number is subtracted from the average conservation number and this value is multiplied by a constant A (i.e. $[CS_i - C_{av}] \times A$). This value is added to the averaged conformational parameters for α-helices and β-sheets in the aligned sequences. The constant A is included to highlight differences in conservation. Optimum values were found to be $A = 150$ for $C_{av} \leq 0{\cdot}55$ and $A = 250$ for $C_{av} > 0{\cdot}55$. Thus in the more conserved sequences, gaps and major variation in chemical type of residue lead to a greater penalty in secondary structure prediction. In addition, a gap was assigned the coil and turn propensities of 0·0, and a helix and sheet propensity as the average of the Gly value between $i-3$ and $i+3$. Thus the helix and sheet-breaking character of Gly is used to penalize further the prediction of secondary structures where there are gaps introduced. Trials showed that a 2% improvement occurred when a gap was scored as a Gly rather than simply a propensity of 0·0.

The conservation plot and the predicted secondary structure can be used to locate the functionally important regions of enzymes. These include residues directly involved in catalysis as well as binding sites. FIRs may consist of one or more sequential residues, and within a family of enzymes are frequently invariant. The definitions are taken from the Brookhaven Data Bank SITE record (Bernstein *et al.*, 1977), or if not available, from the literature. The algorithm developed is: (1) an active segment is predicted within a five-residue segment if n or more residues are invariant, where $n = 4$ if $C_{av} > 0{\cdot}5$, otherwise $n = 3$; (2) an active segment is predicted if in a predicted loop region there is one invariant residue at position i (i.e. $C_i = 1{\cdot}0$) and the smoothed conservation obeys the rules $CS_i > 1{\cdot}50 \times C_{av}$, and $CS_i > 0{\cdot}7$. The first part of the algorithm quantifies the extent of sequence invariance (n) judged against the background of the overall similarity of the aligned sequences (C_{av}). The second part uses the principle that invariant residues in loop regions are more suggestive of FIR than invariant residues

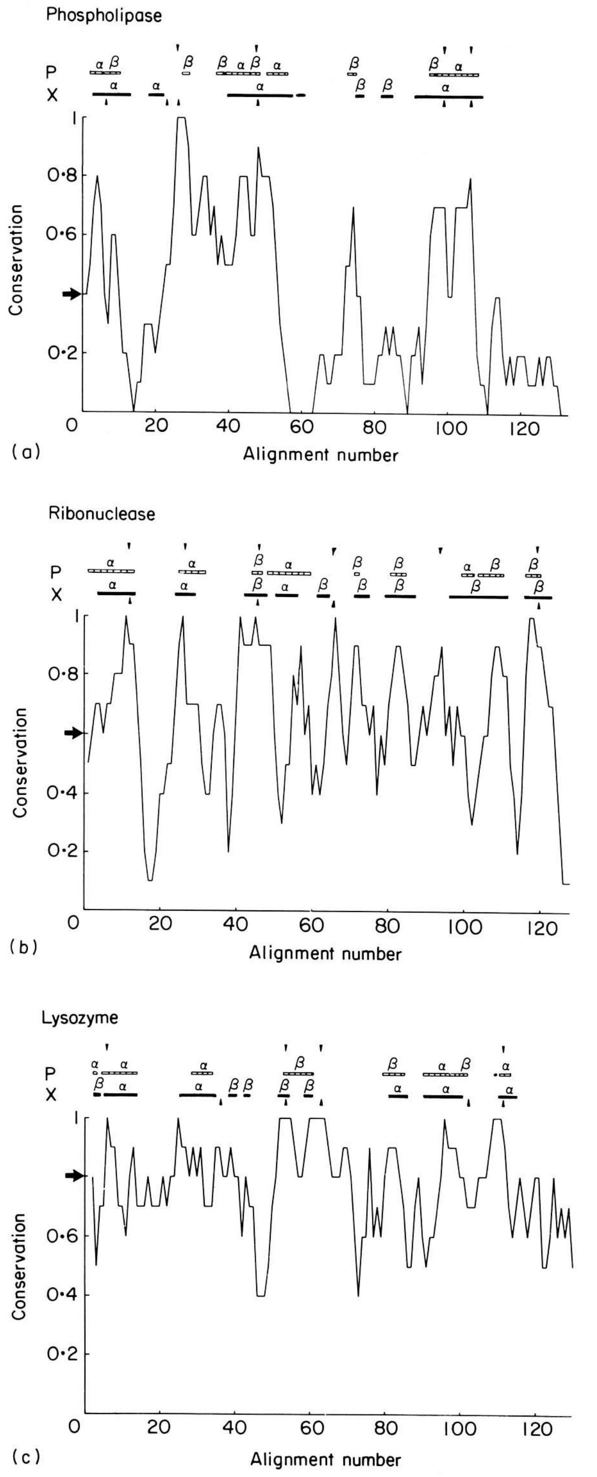

Figure 1. Conservation plots showing predicted and X-ray secondary structure, and active segments. The smoother conservation number (CS_i) is plotted along the sequence. The large arrow denotes the average conservation number for the entire sequence. P and X denoted predicted and X-ray secondary structures. The arrowheads denote predicted and X-ray active segments.

within the secondary structure core. The algorithm is at present based on general principles of FIRs, and a general analysis of the location of FIRs in proteins is an progress, which will then be used to refine the algorithm.

Table 1 gives the accuracies of the predictions of secondary structure of the 11 proteins that were considered, as they had both a known secondary structure and more than four homologous sequences. When the Robson algorithm was applied to just the single sequence, the average accuracy of a three-state prediction (coil and turn being considered as one state) was 57% (Table 1). The alignment of the sequences and the use of an averaged propensity of the sequence and the use of an averaged propensity for all residues at an aligned position yielded an average improvement of 4% to 61%.

Figure 1 illustrates the conservation plot and the secondary structure prediction for three proteins. The minima of the conservation plot generally occur in the loop regions between the secondary structures and thus provide information to help in structure prediction. When the averaged Robson prediction is modified by the conservation plot, the accuracy increases to 66%. This represents a 9% improvement on the prediction on one sequence (Table 1, Fig. 1) and a 5% improvement from the averaged prediction. For each of the 11 proteins, which cover all the structural classes, there is an improvement in prediction. It is important that there is increased accuracy for the $\alpha+\beta$ proteins, which are not amenable to improvement by a template approach. Inspection of the results shows that, as intended, the improvements occur both by the promotion of α and β structure in the regions of high sequence conservation and by penalizing the prediction of these regular secondary structures in sections of low sequence conservation.

Table 3 gives the results of the prediction of the FIRs. The algorithm located 26 out of the 43 FIRs in the seven enzyme families. In addition, only eight segments were overpredicted and of these five included a Cys in a conserved disulphide bridge. The prediction of the FIRs is also illustrated in Figure 1. Table 3 also shows the results of predicting FIRs with cruder algorithms. Simply considering a single identity leads to a high level of overprediction (136 extra FIRs). In contrast, the search for three consecutive identities locates far fewer FIRs than the proposed algorithm.

Plots of sequence conservation or variability have been presented. One major application is the work of Kabat (1976), who plotted sequence variability for the immunoglobulin domains and highlighted the variable loops that form the antigen binding regions. However, this paper presents the use of such plots for general secondary structure prediction and formalizes concepts that were previously incorporated by hand in an unsystematic manner. Further work is required on many aspects of the algorithm. The best method for quantifying sequence variation, including gaps, needs to be determined. An alternative simple approach instead of using the Venn diagram would be to use an average value of the sequence alignment score (Barker *et al.*, 1986). However, an analysis of the types of residue substitutions and

Table 3

Prediction of functionally important residues

Enzyme	No. of FIR segments	1 identity		3 consecutive		Algorithm		No. Cys in extra
		No. located	No. extra	No. located	No. extra	No. located	No. extra	
Kallikrein	3	2	25	2	3	2	3	2
Lactate dehydrogenase	9	8	29	3	1	3	1	0
Triose phosphate isomerase	6	6	26	4	1	5	1	0
Dihydrofolate reductase	10	8	2	1	0	5	0	0
Phospholipase A_2	6	5	8	1	0	4	0	0
Ribonuclease S	4	4	19	4	1	4	2	2
Lysozyme	5	5	27	3	2	3	1	1
Total	43	38	136	18	8	26	8	5

The results of predicting FIRs with different approaches are given; 1 identity, refers to simply loacating 1 invariant residue along the sequence; 3 consecutive, denotes a search for 3 consecutive identities. The results of the proposed algorithm are then given.

gap insertions that occur in the regular secondary structures and in their connections would yield empirical scoring schemes for the conservation value and scaling factor (A) used in this algorithm. This analysis should quantify the level of sequence similarity within the family and relate this to the expected improvement in prediction. If the sequences are too similar then there is little additional information, but if there is widespread sequence variation then the assumption of identical secondary structure no longer holds. Similarly, an analysis of the location of active site residues in proteins would improve the prediction of FIRs.

The aim of this work was to develop an approach for predicting secondary structure and FIRs that is automatic and without any subjective judgement. Furthermore, with their probabilistic nature, the algorithms are flexible enough to cope with some errors in sequence or alignment. The 9% improvement in secondary structure prediction for all structural classes will provide a basis for further improvements incorporating feedback of structural class to modify cut-off parameters (Garnier *et al.*, 1978) and to select suitable templates (Taylor & Thornton, 1983, 1984). The identification of possible FIRs can be used to locate target residues for site-specific mutagenesis.

We thank the SERC for support and Professor T. L. Blundell and Dr Janet Thornton for helpful discussion.

Markéta J. Zvelebil
Geoffrey J. Barton
William R. Taylor
Michael J. E. Sternberg

Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX, U.K.

Received 10 December 1986

References

Barker, W. C., Hunt, L. T., Orcutt, B. C., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Johnson, G. C., Siebel-Ross, E. I. & Ledley, R. S. (1986). Protein Information Resource, National Biomedical Research Foundation, Georgetown University, Washington D.C.

Barton, G. J. & Sternberg, M. J. E. (1987). *Protein Engineering*, **1**, 89–94.

Bernstein, F. C., Koetzle, T., William, G. J. B., Meyer, E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Chou, P. Y. & Fasman, G. D. (1978). *Advan. Enzymol.* **47**, 45–148.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1983). *Biochemistry*, **22**, 4894–4904.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986). *Biochemistry*, **25**, 266–275.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). *J. Mol. Biol.* **120**, 97–120.

Greer, J. (1981). *J. Mol. Biol.* **153**, 1027–1042.

Kabat, E. A. (1976). *Structural Concepts in Immunology and Immunochemistry*, 2nd edit., Holt, Rinehart & Winston, New York.

Kabsch, W. & Sander, C. (1983). *FEBS Letters*, **155**, 179–182.

Lesk, A. M. & Chothia, C. (1980). *J. Mol. Biol.* **136**, 225–270.

Levin, J. M., Robson, B. & Garnier, J. (1986). *FEBS Letters*, **205**, 303–308.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Lim, V. I. (1974). *J. Mol. Biol.* **88**, 873–894.

Murata, M., Richardson, J. S. & Sussman, J. L. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 3073–3077.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.

Nishikawa, K. & Ooi, T. (1986). *Biochim. Biophys. Acta*, **871**, 45–54.

Sternberg, M. J. E. (1986). *Anti-Cancer Drug Design*, **1**, 169–178.

Sweet, R. M. (1986). *Biopolymers*, **25**, 1565–1577.

Taylor, W. R. (1986*a*). *J. Theor. Biol.* **119**, 205–218.

Taylor, W. R. (1986*b*). *J. Theor. Biol.* **188**, 233–258.

Taylor, W. R. (1987). In *Nucleic Acid and Protein Sequence Analysis—A Practical Approach* (Bishop, M. & Rawlings, C., eds), pp. 285–321, IRL Press, Oxford.

Taylor, W. R. & Thornton, J. M. (1983). *Nature (London)*, **301**, 540–542.

Taylor, W. R. & Thornton, J. M. (1984). *J. Mol. Biol.* **173**, 487–514.

Edited by A. Klug