Improvements in a secondary structure prediction method based on a search for local sequence homo|ogies and its use as a model budding tool

Jonathan Levin, Jean Garnier

BBA 33175

Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool

Jonathan M. Levin and Jean Garnier

Laboratoire de Biochimie physique, INRA, Bâtiment 433, Université de Paris Sud, Orsay (France)

(Received 8 March 1988)

Key words: Protein secondary structure; Amino acid sequence; Sequence homology; Prediction; Protein family; Model building

This report describes an optimised version of a secondary structure prediction method based on local homologies, using a new data base. A 63% prediction accuracy, for three states, was obtained after elimination of the protein to be predicted and all proteins with a percentage identity greater than 22% from the data base. This corresponds to a 5% increase in accuracy on the original method (Levin et al. FEBS Lett. 205 (1986) 303–308). The flexibility of the method to the incorporation of information extraneous to the prediction was demonstrated by the prediction of the homologous proteins in the data base. Using the percentage identity with the protein to be predicted, to weight the relative importance of each protein, for all proteins with a percentage identity greater than 30%, the mean correct prediction per chain was 87%. As a result this algorithm can be used during the molecular modelling process, both to give an idea of the structural similarity between two proteins and as an aid in the determination of the best alignment. Incorporation of the result of a protein folding type assignment based on the global amino-acid composition increased the overall prediction to 66%.

Introduction

The ever widening gap between the number of known sequences (putative or otherwise) and the number of known structures present a frustrating bottleneck for all those researchers wishing to know something about the structural and functional aspects of a particular protein. Conformational determination by computer sequence analysis (most often seen in the form of secondary structure prediction methods), is the only way to glean some structural information about a protein, which has no homologous counterpart of known structure, without resorting to lengthy and not always successful crystallographic and NMR studies (with the exception of CD experiments which often contain large errors). Unfortunately, to date the knowledge gained by computer sequence analysis has been rather limited and any improvement in the quality of the information so gained as well as in the degree of confidence one can place in it, must be considered a useful advance. The corollary of this, is the ever increasing interest in modelling a protein on a homologous protein of known structure. Whilst this is fairly straightforward for a pair of related proteins with a percentage identity greater than 50%, below this value, insertions/deletions in the amino-acid sequences, often sig-

Abbreviations: K&S algorithm, algorithm according to Kabsch and Sander (see Ref. 15); MIR, multiple isomorphous replacement; H, helix; E, β strand; C, coil.

Correspondance: J. Garnier, Laboratoire de Biochimie physique, INRA, Bâtiment 433, Université de Paris Sud, 91405 Orsay, France.

naling structural differences, make for a more difficult procedure. The first step in the three-dimensional modelling of one protein on another is the alignment of the amino-acid sequences, in order to determine the correspondance between the residues. For distantly related proteins, alignments based on sequence information alone often give results inconsistent with the alignments obtained by superposition of the three-dimensional structures [1]. A secondary structure prediction algorithm, if sufficiently accurate, i.e., that all the secondary structure elements are correctly predicted, could help to find the best alignment by aligning the secondary structures.

This article presents a revised and updated version of a secondary structure prediction algorithm based on local sequence homology by Levin et al. [2], coupled with a prediction of the folding type based on the global amino-acid composition. This algorithm falls in the category of knowledge-based prediction methods in that it requires a data base of known structures for comparison with a protein to be predicted. The most widely used secondary structure prediction algorithms [3–5] also belong to this category.

The algorithm is based on the hypothesis that similar peptide sequences have similar secondary structure tendencies. This idea has been previously explored [2,6,7]. It is then merely necessary to search in a data base of observed secondary structures to find a sufficiently large number of similar peptides and assign to the unknown peptide the conformation most commonly observed. Similarity is defined by calculating a match score between two peptides, using a similarity matrix (see below), and rejecting those peptides below a certain cutoff. A peptide length significantly greater than 5 is necessary, because, as Kabsch and Sander [8] and Argos [9] pointed out, pentapeptides of identical sequence can adopt completely different conforn. ations.

The percentage of correctly assigned conformations after prediction can be improved by adding in new information. The most obvious example of this is the case of a homologous protein in the data base, the assignment need only then be weighted towards this model structure. However, a prediction of the protein folding type if sufficiently accurate would also improve the percentage of correctly assigned conformations [3]. Klein and Delisi [10] developed a method for determining the protein folding type based on a statistical analysis of four different secondary structure prediction methods. As these methods tend to extract the same type of information from the primary protein structure, i.e., conformational tendencies based on local sequence analysis, it would be more interesting, a priori, to employ a folding-type determination not based on information deduced from previous secondary structure prediction. Nishikawa et al. [11–13] showed that the global amino-acid composition is related to the protein-folding type and Nakashima et al. [14] developed a prediction method based on this. Using an algorithm similar to that of Nakashima et al. [14]. We can improve the overall secondary structure prediction accuracy by weighting the protein to be predicted towards the predicted folding type. The results are also presented with an analysis of the confidence one can place in their accuracy.

Materials and Methods

Any knowledge-based prediction is only as good as the information in the data base. Kabsch and Sander (K&S) [15] defined an objective algorithm for the assignation of secondary structure in proteins. Using this algorithm they published a dictionary of secondary structure using 62 proteins from the Brookhaven data bank [16]. The K&S algorithm uses the presence of certain characteristic hydrogen bonds for the assignment of regular secondary structure. Imprecisions can sometimes be found in coordinates obtained from an initial MIR (multiple isomorphous replacement) phased electron-density map. For example, Sielecki et al. [17] report an overall shift of 0.83 Å during refinement (for main-chain atoms) with some shifts of 3 to 4 Å. These imprecisions can result in an initial model missing the necessary hydrogen bonds used to define the secondary structure. Considerable changes can be seen in the secondary structure assignments before and after refinement. On example of this is the acid (aspartyl) proteinase endothiapepsin, where between versions 2APE and 4APE (Brookhaven code), the number of amino-acid residues assigned by the K&S algorithm as β

TABLE I

THE 67 PROTEINS IN THE DATA BASE

File names and depositors are from the Brookhaven Data Bank (1987)

Protein	File name	Depositors
Acid proteinase (*E. parasitica*)	4APE	T. Blundell
Acid proteinase (*P. janthinellum*)	2APP	A. Sieliki, M. James
Actinidin	2ACT	E. Baker
Agglutinin (wheat germ)	3WGA	C. Wright
Alcohol dehydrogenase (apo)	4ADH	C.I. Branden
α-Lytic proteinase	2ALP	Fujinaga, Delbaere, Brayer, James
Aspartate carbamoyltransferase	4ATC	W. Lipscomb
Azurin (*Alcaligenes denitrificans*)	1AZA	E. Baker, G. Norris
α-Bungarotoxin	2ABX	R. Love, R. Stroud
Ca-binding parvalbumin	1CPV	R. Kretsinger
Ca-binding protein (intestinal)	3ICB	C. Wright
Carbonic anhydrase B (human)	2CAB	K. Kannan
Carboxipeptidase A	5CPA	D. Rees, W. Lipscomb
Catalase (bovine liver)	8CAT	I. Fita, M. Rossmann
α-Chymotrypsin A (bovine)	5CHA	R. Blevins, A. Tulinsky
Citrate synthase (porcine)	2CTS	Remington, Wiegand, Huber
Crambin	1CRN	W. Hendrickson, M. Teeter
γ-II-Crystallin (calf)	1GCR	T. Blundell
Cytochrome *c* (Albacore tuna)	3CYT	T. Takano, R. Dickerson
Cytochrome *c* (rice)	1CCR	H. Ochi, N. Tanaka
Cytochrome *c* (prime)	2CCY	B. Finzel et al.
Cytochrome *c* peroxidase (yeast)	2CYP	B. Finzel, T. Poulos, J. Kraut
Cytochrome c_2 (reduced)	3C2C	G. Bhatia, B. Finzel, J. Kraut
Cytochrome c_3 (*D. vulgaris*)	2CDV	N. Yasuoka, M. Kakudo
Cytochrome *c*-551 (oxidized)	351C	Matsuura, Takano, Dickerson
Dihydrofolate reductase (*L.* casei)	3DFR	Filman, Metthews, Kraut
Elastase (porcine)	2EST	L. Sieker, D. Hughes
Erabutoxin B (sea snake)	2EBX	B. Low
Erythrocruonin (reduced deoxy)	1ECD	Steigemann, Weber
Ferredoxin (*P. Aerogenes*)	1FDX	Adman, Sieker, Jensen
Ferredoxin (*S. Platensis*)	3FXC	M. Kakudo
Falvodoxin (Clos. MP, oxidized)	3FXN	M. Ludwig
Ferredoxin (*Azobacter vinelandii*)	2FD1	C. Stout
Glutathione peroxidase (bovine)	1GP1	O. Epp, R. Ladenstein
Hemerythrin (met)	1HMQ	Stemkamp, Sieker, Jensen
Hemoglobin (human, deoxy)	2HHB	G. Fermi, M. Perutz
Hemoglobin V (cyano, met, lamprey)	2LHB	Honzatko, Hendrickson, Love
High potential iron protein	1HIP	J. Kraut
IGG FAB (kappa) MCPC603	1MCP	Satow, Cohen, Padlan, Davies
Immunoglobulin FAB (Lambda) Kol	1FB4	M. Marquart, R. Huber
Immunoglobulin B–J (V-Dimr)	1REI	O. Epp, R. Huber
Immunoglobulin B–J (V-Mnmr) RHE	2RHE	Furey, Wang, Yoo, Sax
Kallikrein (porcine)	2PKA	W. Bode, Z. Chen
Lactate dehydrogenase	4LDh	W. Eventoff, M. Rossmann
Leghemoglobin	1LH1	Vainshtein, Harutyunyan
Lysozyme (bacteriophage T4)	2LZM	L. Weaver, B. Matthews
Lysozyme (human)	1LZ1	P. Artimiuk, C. Blake
Melittin	1MLT	T. Terwilliger, D. Eisenberg
Myoglobin (sperm whale, met)	1MBN	H. Watson
Scorpion Neurotoxin (variant)	1SN3	C. Bugg et al.
Ovomucoid third domain (quail)	1OVO	W. Bode, O. Epp
Papain D	1PPD	J. Jansonius

TABLE I (continued)

Protein	File name	Depositors
Phospholipase A2 (bovine)	1BP2	B. Dijkstra, J. Drenth
Plastocyanin	1PCY	J. Guss, H. Freeman
Prealbumin (human plasma)	2PAB	S. Oatley, C. Blake
Proteinase A (*S. griseus*)	2SGA	M. James, A. Sielecki
Proteinase II (rat mast cell)	3RP2	S. Remington, B. Matthews
Ribonuclease A	1RN3	Borkakoti, Moss, Palmer
Rubredoxin	5RXN	K. Waterpaugh
Staphylococcal nuclease	2SNS	Legg, Cotton, Hazen
Subtilisin BPN prime	1SBT	J. Kraut
Superoxide dismutase	2SOD	J. Richardson, D. Richardson
Thermolysin	3TLN	B. Matthews, M. Holmes
Trypsin (orthorhombic)	1TPO	W. Bode, J. Walter, R. Huber
Trypsin inhibitor (bovine)	4PTI	R. Huber, J. Deisenhofer
Virus (satellite tobacco necrs)	2STV	T.A. Jones, L. Liljas
Virus coat protein (SBMV, $T = 3$)	4SBV	M. Rossman

strand went up from 33 to 50%. 2APE is at 2.5 Å resolution and partially refined and 4APE is at 2.1 Å with a crystallographic R value of 0.158.

To avoid discrepencies and inconsistencies in the data base, the K&S algorithm was applied to 67 well refined proteins (see Table I) at high resolution (resolution greater than 2.8 Å with a crystallographic R factor of less than 0.25) with known sequences from the Brookhaven data bank. As mentioned above the K&S algorithm is very sensitive to the coordinates of the main-chain atoms even for well refined proteins, thus slight variations in the assignments of secondary structure can be seen between identical chains in oligomeric proteins. This is particularly noticable for turn assignments which include sections of distorted helices. To reduce these differences and thus create a more coherent data base, the 8 state assignment of K&S was reduced to three states: H (= helix), defined as all amino-acid residues assigned as H, runs of 4 or more Gs and runs of 3 Gs next to a run of Hs; E (= β strand), all amino-acid residues assigned as E; C (= coil), all amino-acid residues not H or E. There are 12058 amino-acid residues in the data base, 27% helix, 22% β strand, and 51% coil.

As the prediction algorithm is based on a search for homologous peptides, it is imperative to use a data base free of homologous proteins: for an accurate assessment of the predictive capabilities of the algorithm. A sequence-matching program was written to determine the best match between any two amino-acid sequences, allowing for gaps (insertions and deletions). The algorithm used was a slightly modified version of the algorithm developed by Needleman and Wunsch [18]. The matrix used to calculate the match scores was the identity matrix. Using this program the percentage identities between each polypeptide chain in the data base were calculated. This percentage is defined as

TABLE II

PRINCIPLE HOMOLOGIES IN THE DATA BASE

For file name see Table I.

File name	Identity (%)	File name	File name	Identity (%)	File name
4APE	55	2APP	2ACT	48	1PPD
2ALP	39	2SGA	2ABX	42	2EBX
5CHA(1)[a]	42	2EST	5CHA(1)	34	2PKA(1)
5CHA(1)	43	1TPO	5CHA(1)	37	3RP2
5CHA(2)	43	2EST	5CHA(2)	41	2PKA(2)
5CHA(2)	45	1TPO	2EST	42	2PKA(1)
2EST	30	2PKA(2)	2EST	42	1TPO
3RP2	36	1TPO	3CYT	59	1CCR
3CYT	44	3C2C	1CCR	40	3C2C
3RP2	38	2EST	1FDX	43	2FD1
2HHB(1)	45	2HHB(2)	2HHB(1)	35	2LHB
1MCP(1)	64	1REI	1MCP(1)	47	2RHE
1MCP(2)	44	1FBA(2)	1REI	51	2RHE
2PKA(1)	38	3RP2	2PKA(1)	38	1TPO
2PKA(2)	34	3RP2	2PKA(2)	41	1TPO

[a] The numbers in brackets refer to the polypeptide chains.

G	2																			
P	0	3																		
D	0	0	2																	
E	0	-1	1	2																
A	0	-1	0	1	2															
N	0	0	1	0	0	3														
Q	0	0	0	1	0	1	2													
S	0	0	0	0	1	0	0	2												
T	0	0	0	0	0	0	0	0	2											
K	0	0	0	0	0	1	0	0	0	2										
R	0	0	0	0	0	0	0	0	0	1	2									
H	0	0	0	0	0	0	0	0	0	0	0	2								
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2							
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2						
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2					
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2				
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2			
F	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	0	-1	0	2			
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	-1	0	1	2		
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	-1	0	0	0	2
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W

Fig. 1. The secondary structure similarity matrix (Levin et al. Ref. 1) which gives a score for the replacement of one amino acid by another. Reprinted with kind permission from FEBS Letters.

the number of identical pairs of amino-acid residues after the best match has been calculated divided by the number of amino-acid residues in the smaller of the two sequences multiplied by 100. The data base contains 72 polypeptide chains, 28 of which have a percentage identity greater than or equal to 30% (see Table II).

The prediction algorithm is essentially the same as that which was published in Ref. 1. The first part of the algorithm, which consists of a search for homologous peptides, uses the similarity matrix previously published (Levin et al., Ref. 1) (see Fig. 1). The algorithm makes a comparison between every sequence of n amino-acid residues, where n is the window length, in the test protein with every fragment of length n amino-acid residues in the data base. If the calculated match-score between the two peptides is less than a cutoff value the peptide is rejected. Every time a peptide is found whose score is greater than or equal to the cutoff value, its observed conformation is assigned to the test sequence with its similarity score (see Table III). Once every fragment in the test protein has been compared, the secondary structure attributed to each residue is that which has the highest value after multiplication by the following decision constants: DCH = 1.067, DCE = 1.25, and DCC = 0.75, for the H, E and C conformations, respectively (see Table IV).

Using this data base and these decision constants the optimal values of the window size and cut off were determined by an exhaustive series of runs. The protein to be predicted and all proteins with a percentage identity greater than 22% were removed from the data base for each prediction run.

In order to determine the protein folding type the proteins in the data base were divided into four classes: α-rich, greater than 35% helix and less than or equal to 10% β strand; β-rich, greater than 30% β strand and less than or equal to 10%

TABLE III

AN EXAMPLE OF SECONDARY STRUCTURE ASSIGNMENTS

In the above example the window length (n) is 7 with a cutoff of 7. Three homologous fragments were found for amino-acid residues 1–7 of the test sequence, i.e., those whose similarity scores were greater than or equal to the cutoff. The first with a score of 7 had an observed conformation of CHHHHHC, the second had a score of 7 and an observed conformation of CCEEEEC, and the score for the third was 8 with a conformation of CCHHHHC. Then the first amino-acid residue of the test sequence (column res) is credited with 7 + 7 + 8 in column C which is the only observed conformation of the three fragments for the first amino-acid residue. The second amino acid of the test sequence is credited with the score of 7 in column H which is the observed conformation of the second amino-acid residue of the homologous fragment with a score of 7 and a score of 7 + 8 in column C, which is the observed conformation of the other two homologous fragments at this position and so on for the rest of the test sequence. Two homologous fragments were found for amino-acid residues 2–8, each had a score of 9 and a conformation of CHHHHCC AND CCEEEEC and their scores were added to the above table following the procedure described above for amino-acid residues 2–8. The prediction is then optimised using the decision constants. The sums in column H are multiplied by 1.067, those in column E by 1.25, and those in column C by 0.75. The prediction is then based on the conformation with the highest score, so for residues 1–8 it is CCHHHHCC (see Table IV).

Res	Conformation		
	H	E	C
1			7+7+8
2	7		7+8+9+9
3	7+8+9	7	9
4	7+8+9	7+9	
5	7+8+9	7+9	
6	7+8+9	7+9	
7		9	7+7+8+9
8			9+9

TABLE IV

AN EXAMPLE OF SCORING VALUES FOR THE SECONDARY STRUCTURE PREDICTION

The values are for the example given in Table III. The predicted conformation for each residue, 1 to 8, is listed in column 5 according to the highest score of the three conformations in columns 2–4.

Res	Conformation			Pred
	H	E	C	
1	0	0	17	C
2	7	0	25	C
3	26	9	7	H
4	26	20	0	H
5	26	20	0	H
6	26	20	0	H
7	0	11	23	C
8	0	0	14	C

helix; low secondary structure, regular secondary structure (helix and β strand) less than 25%; mixed, all polypeptide chains not in the other three categories. The average (%) content for each amino acid and standard deviations were calculated for each class (see Table III). To determine the folding type, the closeness (C) of the global amino-acid composition to each of the four protein classes (j) was calculated:

$$C_j = \sum_{i=1}^{20} (\mathrm{abs}(\mathrm{Av}_{ij} - \mathrm{Per}_i)/\mathrm{Sd}_{ij}) W_i$$

Per is the (%) content of amino acid (i) in the unknown protein, Av and Sd are the class average (%) content for each amino acid and standard deviations, respectively. C was calculated for each of the protein classes and the class predicted was the protein class associated with the smallest value of C (the C for the low secondary structure class was multiplied by a factor of 1.3). W is an optimised weighting factor for each amino acid, independent of class type (see Table V).

Results

When the protein to be predicted and all other proteins with a percentage identity greater than

TABLE V

THE AVERAGE % AMINO-ACID CONTENT (Av) AND STANDARD DEVIATIONS (Sd) FOR EACH OF THE FOUR CLASSES

For definitions of class type see text. There are 20 α-rich, 20 β-rich, 25 mixed and 7 low secondary structure proteins.

Amino acid	α-Rich		β-Rich		Mixed		Low SS		W
	Av	Sd	Av	Sd	Av	Sd	Av	Sd	
GLY	7.5	2.3	10.2	3.2	8.4	3.2	9.5	7.4	0.45
PRO	4.0	1.7	4.6	1.6	4.8	2.3	7.0	3.3	0.33
ASP	5.9	3.0	4.4	2.0	5.4	2.0	8.9	5.5	1.04
GLU	6.3	3.8	4.0	2.3	4.7	2.6	6.9	4.0	1.03
ALA	11.5	5.2	6.5	3.1	8.2	2.4	9.3	7.0	1.66
ASN	3.5	1.9	4.9	2.4	5.8	2.3	3.9	1.7	1.05
GLN	3.5	2.0	4.6	2.5	3.3	1.5	3.3	2.3	1.32
SER	5.2	1.9	11.1	3.5	6.8	2.9	5.3	3.8	0.56
THR	5.3	1.6	7.8	3.0	6.4	2.4	5.3	4.3	1.52
LYS	10.2	3.0	4.1	2.1	7.1	4.0	4.8	2.6	0.41
ARG	2.4	1.9	3.4	2.7	3.7	2.8	1.5	1.5	0.56
HIS	2.9	2.5	2.2	1.7	2.1	1.9	1.0	1.0	0.65
VAL	6.5	2.7	7.3	2.6	6.8	2.6	5.6	3.2	1.32
ILE	4.2	2.9	5.1	1.6	4.7	2.7	5.2	4.1	1.05
MET	1.6	1.1	1.0	1.1	1.9	1.3	0.9	0.7	0.49
CYS	0.9	0.8	3.4	3.4	4.0	4.1	10.6	5.2	1.50
LEU	9.6	3.5	6.2	1.8	6.9	2.7	3.8	2.7	0.82
PHE	4.9	2.6	3.8	1.8	3.2	1.7	2.0	1.7	1.07
TYR	2.4	1.8	3.8	1.9	4.3	2.2	3.6	1.8	0.59
TRP	1.6	1.0	1.6	0.9	1.4	1.1	1.4	1.2	0.36

TABLE VI

PREDICTION RESULTS

The protein to be predicted and all proteins with a percentage identity greater than 22% were excluded from the calculation. Results are obtained with the new data base (Table I) and are given for a three state prediction (helix, β strand and coil) as the number of correctly predicted residues divided by the total number of residues multiplied by 100.

Window length	Cutoff				
	3	4	5	6	7
7	60.1	61.0	61.1	61.0	59.8
8		61.5	61.6	61.5	61.0

22% were removed from the data base, the algorithm correctly assigned 59.8% of the amino-acid residues (number of amino-acid residues correctly assigned divided by the total number of amino-acid residues) for a three-state prediction (helix, β strand and coil), with a window length of 7 and a cutoff of 6 and using the data base published by K&S. With the same parameters but using the new data base 61.0% of the amino-acid residues were correctly assigned.

Optimisation of the prediction

We wished to determine the optimum window length and cutoff. To this end a series of tests were performed using the new data base, at varying window lengths and cutoffs (see Table VI and Fig. 2). For short window lengths, the optimal cutoff was 5, for longer window lengths the optimal cutoff was 7. The best overall prediction, 63%, was obtained with a window length of 17 and a cutoff of 7 (see Fig. 2). The mean correct prediction per polypeptide chain was 64.4% with a standard deviation of 8.95%, the best prediction was 95% and the worst 46% (see Table VII column B). Although with the original data base the proportions of each of the three states were respected, this proved difficult with the new data base. The helix content was correctly predicted, but there was a 26% under prediction of β strand. An increase in the decision constant for β strands results in a decrease in the overall number of correctly assigned amino-acid residues. This can be explained by looking at the class of proteins observed as β-rich and the class of proteins observed as mixed. There is an under prediction of β strand of 10% for mixed type proteins but 48% for β-rich proteins. An increase in the decision con-

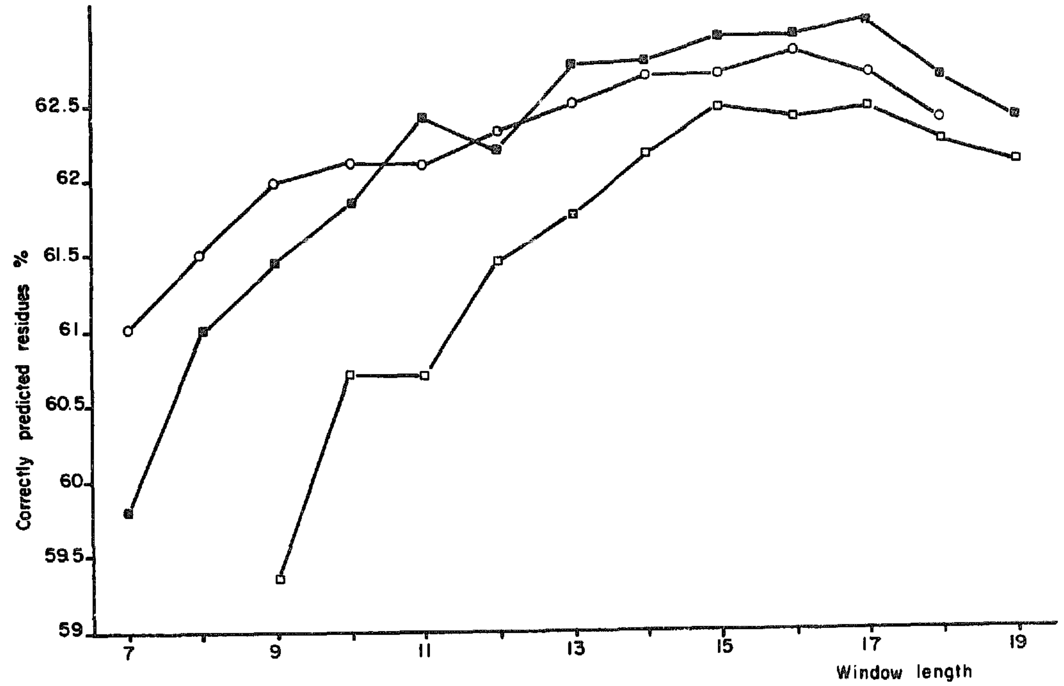

Fig. 2. Prediction accuracy as a function of window length for three different cutoff values. The optimal conditions corresponding to the best prediction are a window length of 17 and a cutoff of 7. ○, cutoff of 6; ▨, cutoff of 7; □, cutoff of 8.

TABLE VII

COMPARISON OF THE RESULTS FOR EACH PROTEIN OF THE BASIC PREDICTION (B), HOMOLOGY PREDICTION (H) AND FOLDING TYPE ASSIGNMENT AND PREDICTION (T)

Window length 17 and cutoff 7 for all three predictions. B is the basic prediction where all proteins with a percentage identity greater than 22% were removed. H is the prediction explicity using the homologies in the data base. The scores of the similar peptides from each protein in the data base are multiplied by its % identity with the protein to be predicted, if this % identity is greater than 30% the scores are multiplied by 2 as well (see text for detailed explanation). T is the prediction using the protein folding type assignment with the homologous proteins excluded from the prediction as in B. The assignments are given as one of for classes: α, α-rich; β, β-rich; M, mixed type; L, low secondary structure. The observed and predicted folding types are given with and without the weighting factor W of Table V. The type assignment used in column T includes W. If the assignment is for low secondary structure or mixed type proteins the proteins of this class observed in the data base have their scores increased by 10%. If the assignment is for α-rich proteins, the scores for the α-rich proteins in the data base are increased by 10–150% depending on the difference between the value of C_{helix} and the next lowest value of C, idem for β-rich proteins, except the minimum increase is 40% (see text for detailed explanation).

File Name	B (%)	H (%)	Obs	Type assignment	Type assignment + W	T (%)
4APE	57 6	88.8	β	β	β	64.2
2APP	56.4	89.5	β	β	β	64.4
2ACT	53.7	83.5	M	M	M	52.8
3WGA	74.1	72.9	L	L	L	74.7
4ADH	54.8	56.2	M	M	M	54.8
2ALP	50.0	77.8	β	β	β	67.7
4ATC(1)	59.7	59.4	M	α	α	59.4
4ATC(2)	55.7	53.6	M	M	M	56.2
1AZA	45.7	46.5	M	M	M	45.7
2ABX	81.1	75.7	L	M	L	81.1
1CPV	58.3	62.0	α	α	α	57.4
3ICB	94.7	89.3	α	α	α	78.7
2CAB	64.8	62.5	M	M	M	64.5
5CPA	69.7	67.4	M	β	β	63.5
8CAT	65.3	65.7	M	M	M	65.7
5CHA(1)	65.7	89.3	β	β	β	69.5
5CHA(2)	60.8	87.6	M	M	M	60.8
2CTS	65.2	69.6	α	M	α	67.7
1CRN	56.5	47.8	α	L	L	58.7
1GCR	54.6	57.5	β	M	β	58.1
3CYT	57.3	99.0	α	M	M	57.3
1CCR	61.3	91.9	α	M	M	61.3
2CCY	71.7	74.8	α	α	α	90.6
2CYP	66.6	68.6	α	M	α	68.3
3C2C	64.3	87.5	α	α	α	62.5
2CDV	73.8	70.1	M	M	M	73.8
351C	80.5	78.0	α	α	α	80.5
3DFR	61.7	58.6	M	α	α	59.9
2EST	62.1	86.7	β	β	β	68.3
2EBX	71.0	96.7	β	β	β	71.0
1ECD	74.3	80.1	α	α	α	82.4
1FDX	70.4	75.9	L	L	L	70.4
3FXC	67.4	69.4	L	L	L	68.4
3FXN	68.1	66.7	M	M	M	68.1
2FD1	64.2	70.8	L	L	L	64.2
1GP1	62.5	63.0	M	M	M	62.5
1HMQ	52.2	52.2	α	α	α	56.6
2HHB(1)	61.7	92.9	α	α	α	82.3
2HHB(2)	56.9	89.7	α	α	α	81.5

TABLE VII (continued)

File Name	B (%)	H (%)	Obs	Type assignment	Type assignment + W	T (%)
2LHB	73.2	88.6	α	α	α	81.9
1HIP	56.5	57.7	L	L	L	58.8
1MCP(1)	67.3	83.2	β	β	β	74.1
1MCP(2)	67.6	88.7	β	β	β	72.1
1FB4(2)	75.6	87.3	β	β	β	77.7
1REI	63.6	94.4	β	β	β	67.3
2RHE	64.9	85.1	β	β	β	70.2
2PKA(1)	63.8	96.3	β	β	β	71.3
2PKA(2)	60.5	91.5	M	M	M	59.2
4LDH	52.9	54.7	M	α	M	52.9
1LH1	76.5	79.8	α	α	α	86.9
2LZM	69.5	69.5	α	M	M	67.7
1LZ1	62.3	63.1	M	M	M	63.9
1MBN	80.4	88.9	α	α	α	87.6
1MLT	76.9	80.1	α	α	α	84.6
1SN3	73.9	70.8	M	M	L	72.3
1OVO	60.7	64.3	M	M	M	60.7
1PPD	63.7	92.0	M	β	β	59.9
1BP2	50.4	52.0	α	M	M	50.4
1PCY	72.7	72.7	β	M	β	75.7
2PAB	55.3	56.1	β	β	β	64.9
2SGA	54.1	76.8	β	β	β	63.0
3RP2	56.7	84.8	β	M	M	57.6
1RN3	60.5	62.1	M	M	M	60.5
5RXN	74.1	70.4	L	L	L	74.1
2SNS	56.7	56.0	M	M	M	56.0
1SBT	65.8	62.9	M	β	β	62.9
2SOD	73.5	71.5	β	M	β	74.2
3TLN	61.4	58.2	M	M	M	59.8
1TPO	61.4	89.7	β	β	β	70.9
4PTI	79.3	74.1	M	M	M	79.3
2STV	57.1	57.6	β	M	β	64.1
4SBV	51.8	50.5	M	M	M	50.9
Global	63.0 [a]	72.8 [a]		74.0%	83.6%	65.7 [a]

[a] Percentage of correctly predicted amino-acid residues for the whole data base (12058 residues) and for the three states.

TABLE VIII

NUMBER OF OCCURRENCE AND PROBABILITY OF CORRECT PREDICTION FOR EACH CONFORMATION AND SCALE VALUE

Confidence scale	conformation							
	H		E		C		total	
	(No.)	(%)	(No.)	(%)	(No.)	(%)	(No.)	(%)
1	946	44	558	45	1137	55	2641	49
2	816	51	518	52	1238	55	2572	53
3	664	64	395	52	1295	67	2354	63
4	448	74	276	61	1268	72	1992	71
5	237	79	169	68	1045	80	1451	78
6	76	86	72	83	900	87	1048	87
Total	3187	58	1988	54	6883	68	12058	63

stant for β strand results in a far larger number of incorrect assignments for β strand in mixed type and α-rich proteins (45 proteins in total) than now correct assignments in the class of β-rich proteins (20 proteins).

Confidence scale

We were also interested in determining the probability of a correct prediction on a residue by residue basis. Thus we defined an empirical confidence scale based on the difference in scores between the best and second best choice of the predicted conformation for each residue. A low scale value means the scores between the best and second-best conformation are very close therefore the choice of conformation is uncertain, a high value means the residue shows a clear preference for a particular conformation. The scale was divided up into six intervals, the probability of a correct prediction increases with the confidence scale value. Table VIII shows the number of amino-acid residues predicted in each interval and the fraction of those residues correctly predicted. The results are shown for each of the three states singly and together.

Use of homology

As shown above, this algorithm has a prediction accuracy equivalent to that of Gibrat et al. [4] when predicting the conformation of a protein with no homologous counterpart in the data base. However, when explicitly using the homology existent between the protein to be predicted and a protein in the data base, the prediction accuracy considerably improves. When the homologous proteins where left in the data base the overall prediction increased from 63 to 67%. Then to further weight the effects of the homologous proteins the scores of the similar peptides from each protein were multiplied by the percentage identity of that protein with the protein to be predicted. This gave a prediction of 72% for the whole data base, this was improved to 73% by multiplying again by 2 if the percentage identity was greater than 30%. For example, if the protein in the data base has a 15% identity with the protein to be predicted, all the matchscores are multiplied by 15, if the identity was 35%, the matchscores would be multiplied by 70. All proteins with a percentage identity greater than 30% with another protein in the data base were predicted at greater than 70% (see Table VII). The average increase for this group of proteins was 24% with a mean correct prediction per polypeptide chain of 87.2%.

Use of the protein folding type assignment

As mentioned above, the β strand content of the class of β-rich proteins is severely under predicted; however, this could be rectified with a sufficiently accurate prediction of the protein folding type. The prediction of the protein folding type correctly assigned 74% of the proteins in the data base. This was increased to 84% by optimisation (see Table VII). This was then used to bias the secondary structure prediction towards the assigned protein folding type. If the assignment was for mixed or low secondary structure type, the scores of the proteins in the data base of this class were increased by 10%. Even when using the observed folding types no improvement was found in the secondary structure prediction by further favouring these two classes of proteins. There were no folding type assignments where β-rich proteins were assigned as α-rich or vice versa; however, if following an incorrect assignment of a mixed or low secondary structure protein into either of these two classes the secondary structure prediction becomes too strongly biased towards that particular class, the prediction accuracy may decrease by as much as 25%. This is particularly noticeable for proteins assigned as α-rich as the overall helix content is already correctly predicted. As a result a sliding scale was developed to progressively increase the bias towards either α- or β-rich proteins according to the difference in the value of C for the class assigned and the next lowest value of C, i.e., the bigger the difference the larger the bias. If the assignment is for an α-rich protein, the scores from the α-rich proteins in the data base are multiplied by F with $F = (C_x - C_h)/1.4$. C_x and C_h are the second smallest and smallest values of C, respectively. The value of F may not exceed the upper and lower limits of 2.5 and 1.1. For a β-rich assignment the scores for β-rich proteins in the data base are multiplied by F with $F = (C_x - C_e)/1.3$. C_x and C_e are the second smallest and smallest values of C, respectively. The value of F cannot exceed the upper and lower limits of 2.5

and 1.4, the higher minimum value of F is because β strand is generally underpredicted.

For the whole data base a 2.7% increase was observed, bringing the overall prediction accuracy from 63.0% to 65.7% (see Table VII for detailed results). The underprediction of β strand decreased from 26 to 9%, the total helix content was unchanged.

Discussion

Secondary structure prediction algorithms are generally sensitive to the data base used. In this study the quality of the information in the data base was improved from that previously used, and the number of proteins extended by a careful choice of the coordinate sets available at the Brookhaven data bank and the subsequent assignment of secondary structure. This led to an increase in the prediction accuracy from 59.8% to 61.0%. By using a program to calculate the best match between any two sequences, we were able to determine the effects of homologous proteins on the prediction accuracy. We were thus able to determine the best choice of window length and cutoff. If is worth noting that the optimal peptide length is 17, the same found by Robson and Suzuki [19] and this appears to imply that 17 amino-acid residues is the limit for local sequence effects.

For an unknown protein with no homologous counterpart in the data base, i.e., a percentage identity less than or equal to 22% with respect to all the other proteins, one could expect 63% of the amino-acid residues to be correctly predicted which is identical to the expectancy obtain by the method GOR, Gibrat et al. [4] and is an improvement of 5% on the previously published version of this algorithm [2], the 62% accuracy obtained in the original version contained the homologous proteins in the data base and thus compares with the 67% obtained by the new algorithm when the homologous proteins are left in the data base. The 63% prediction accuracy obtained compares favourably with the results of Sweet [6] 59% and Nishikawa and Ooi [7] 60%. Gibrat et al. [4] showed that there is only a 5% chance due to statistical variation, assuming a normal distribution, of a 1% increase in the prediction accuracy. However, as large variations are observed in the predictions of the homology algorithm (from 46 to 95%) for each protein we wished to provide an estimate on an amino-acid residue by residue basis of the probability of a correct prediction. The scale developed shows those amino-acid residues with a probability of a correct prediction of approx. 45%, to those with an approx. 85% probability. This scale is printed out along with the sequence and prediction and permits identification of those zones in the sequence which are more likely to be correct. A comparison with the equivalent scale developed by Gibrat et al. [4] shows that the agreement between the two methods is very good. Moreover, this agreement extends to incorrect assignment of secondary structure, suggesting that both methods are picking up the same information, or, conversely, that neither of these two methods are capable of determining those zones in the sequence where the secondary structure is not determined by the local sequence. As a result a secondary structure prediction based on a combination of these two methods improves the overall prediction by only 2–3% instead of a hoped for larger increase given the different nature of the two methods (Biou et al. [20]).

The underprediction of β strand can perhaps be explained by the fact that it is the secondary structure most affected by long-range interactions It remains to be seen if the division of β strand into parallel and antiparallel will improve the discrimination.

Comparison of the columns H and B of Table VII shows the effects of the homologous proteins in the data base. The average increase of the prediction accuracy per chain is 24% for the proteins listed in Table II. Thus, the algorithm is particularly suited to making a secondary structure prediction with a partially homologous protein in the data base, as it is very easy to single out a particular protein or proteins in the data base and increase the weight of the contribution of that protein to the prediction. Moreover, this increased contribution only applies to those parts of the sequence where there is a homology as the peptides whose matchscores are less than the cutoff are rejected, and thus the nonhomologous parts of the sequence are not considered. This algorithm should be very useful for the initial modelisation of a protein homologous to another protein of known structure.

How can a secondary structure prediction aid in the modelisation of a protein? In order to determine the correspondance between the amino-acid sequences, it is necessary to align the sequences. Whilst this generally poses no problem for closely related proteins, for distantly related proteins the correlation between sequence and structural homology is not always obvious, Argos [21], and thus one can obtain several different plausible alignments by using different alignment programs and by varying the input parameters, such as the gap penalties or match matrices used. This raises the problem of how to choose between them. The simplest way to test an alignment is to see whether the secondary structures of the two proteins are well matched. In the event that the conformation of only one of the proteins is known, a secondary structure prediction could be used to replace the observed conformation for the other protein if it was sufficiently accurate. As shown above, the secondary structure prediction algorithm presented here is very reliable when predicting the conformation of a protein with a homologous counterpart of known structure.

This algorithm is in fact particularly well suited to aid the initial modelling procedure. As those zones along the sequence which have no sequence similarity with the homologous protein might well be assigned to a secondary structure different from that observed in the homologous protein, differences between observed and predicted conformation can be indications that there is no spatial correspondence between these residues in the two proteins. Furthermore, should these differences extend over large regions of the proteins then perhaps the modelling procedure is unlikely to lead to a good structure for the unknown protein. Thus the algorithm can aid not only in the alignment of the two sequences but can also give an indication of the overall three-dimensional similarity of the proteins.

One of the principle advantages of this algorithm is that there is no set of overall parameters derived from the data base which is used to make the prediction, thus one can easily alter the importance of a protein or group of proteins with respect to the others. One example of this is the case of a homologous protein mentioned above. Another example is the division of the data base into four folding types and subsequently weighting the proteins in the data base belonging to the same folding type as the protein to be predicted. Obviously a method for determining the folding type of an unknown protein is necessary. The method for the assignment of folding type as presented above would almost certainly not give an 84% correct assignment for proteins not included in the learning data base, as the method was optimised for that data base and the data base includes a series of homologous proteins. A meaningful comparison is very difficult with the work of Nakashimi et al. [14] as they used the crystallographers' assignments for the secondary structure content and they used different criteria for defining the folding class types. However, the assignments we used demonstrate that a significant improvement in the secondary structure prediction accuracy can be obtained even by a less than 100% correct assignment of folding types.

The data base used is sufficiently large and varied to justify the claim that the accuracy of prediction for proteins not in the data base, will be equal to the results presented above. However, the folding type assignment, as a result of the weighting factor, *W*, is likely to be more biased towards data base. A folding type assignment method which is less dependent on the data base is currently under development in our laboratory. Without a large increase in the size of the data base it is unlikely that the figure of 63% will be improved for non homologous proteins. However, when the algorithm is used in conjunction with other information, e.g., another prediction algorithm or the prediction of folding type the prospects for improvement are optimistic. A recent article by Zvelebil et al. [22] improved by 9% an earlier version of the secondary structure prediction method GOR (Garnier et al. [3]) when dealing with the specific case of a family of homologous proteins. We do not think that the limits in secondary structure prediction have yet been reached and therefore the prediction accuracy will continue to rise, albeit by small increments.

A prediction program written in Fortran 77 combining the GOR method and this method has been developed, Biou et al. [20] and is available from the authors.

Acknowledgements

J.M.L. would like to thank the Ministère de la Recherche et de la Technologie for the doctoral studentship they awarded him. Computing was performed at Paris Sud Informatique under a grant from the Ministère de l'Education Nationale. This work was also partially funded by the Fondation pour la Recherche Médicale.

References

1 Lesk, A.M., Levitt, M. and Chothia, C. (1986) Protein Eng. 1, 77–78.

2 Levin, J.M., Robson, B. and Garnier, J. (1986) FEBS Lett. 205, 303–308.

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

4 Gibrat, J.-F., Garnier, J. and Robson, B. (1987) J. Mol. Biol. 198, 425–443.

5 Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.

6 Sweet, R.M. (1986) Biopolymers 25, 1566–1577.

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

8 Kabsch, W. and Sander, C. (1984) Proc. Natl. Acad. Sci. USA 81, 1075–1078.

9 Argos, P. (1987) J Mol. Biol. 197, 331–348.

10 Klein, P. and Delisi, C. (1986) Biopolymers 25, 1659–1672.

11 Nishikawa, K. and Ooi, T.J. (1982) Biochemistry 91, 1821–1824.

12 Nishikawa, K., Kubota, Y. and Ooi, T. (1983) 1 J. Biochem. 94, 981–985.

13 Nishikawa, K., Kubota, Y. and Ooi, T. (1983) J. Biochem. 94, 997–1007.

14 Nakashima, H., Nishikawa, K. and Ooi, T. (1986) J. Biochem. 99, 153–162.

15 Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2673.

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

17 Sielecki, A.R., Hendrickson, W., Broughton, C.G., Delbaere, L.T.J., Brayer, G.D. and James, M.N.G. (1979) J. Mol. Biol. 134, 781–804.

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

19 Robson, B. and Suzuki, E. (1976) J. Mol. Biol. 107, 327–356.

20 Biou, V., Gibrat, J.F., Levin, J.M., Robson, B. and Garnier, J. (1988) Protein Eng., in press.

21 Argos, P. (1987) J. Mol. Biol. 193, 385–396.

22 Zvelebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E. (1987) J. Mol. Biol. 195, 957–961.