Volume 280, number 1, 141–146 FEBS 09476 March 1991

ADONIS 001457939100186H

An extension of secondary structure prediction towards the prediction of tertiary structure

Richard C. Garratt*, Janet M. Thornton and Willie R. Taylor**

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

Received 25 October 1990; revised version received 15 January 1991

Secondary structure prediction parameters and optimised decision constants for use with the method of Garnier et al. ((1978) J. Mol. Biol. 120, 97–120) have been derived for two new and distinct substates of β-structure. These we term *internal* and *external* on the basis of their hydrogen bonding patterns. The profiles of the amino acids for several of the parameters are considerably different in the two substates. Predictions using the new parameters attempt to distinguish the strands at the core of the β-sheet from those at its edges and so restrict the possible topologies in tertiary structure prediction. The potential application of these parameters is illustrated for the class of β/α proteins.

Secondary structure prediction; Tertiary structure prediction; β-Sheet residue; Internal β-residue; External β-residue

1. INTRODUCTION

Over recent years it has become increasingly apparent that our ability to successfully predict the secondary structure of a protein from its amino acid sequence will require an understanding of the long-range interactions that influence protein folding. At present the most commonly used statistical algorithms for the prediction of secondary structure on the basis of sequentially local information fail to achieve an accuracy of better than 65% and provide only limited tertiary structural information [1–3].

We have previously defined two substates of β-structure (*internal* and *external*) on the basis of their distinct hydrogen bonding patterns [4]. The distribution of amino acid types in these two substates was shown to be different and the *external* residues (principally located in the edge strands) were significantly less well predicted than the *internal* residues. We would therefore anticipate that a prediction which distinguished these substates might not only improve the quality of the prediction but also provide tertiary structural information by identifying core strands and edge strands separately. This paper reports the extension of the predictive method of Garnier et al. [5], henceforth termed the GOR method [3] to include these new secondary substates.

Correspondence at current address: J. Thornton, Biomolecular Structure and Modelling Unit, Department of Biochemistry, University College, Gower Street, London, WC1E 6BT, UK

**Present address:* Instituto de Física e Química de São Carlos, Departamento de Física e Ciências dos Materiais, Universidade de São Paulo, Caixa Postal 369, 13560 - São Carlos - SP, Brasil

***Present address:* Laboratory of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK

2. EXPERIMENTAL

2.1. Internal *and* external *residues*

Secondary structural definitions for proteins of known structure were obtained from the dictionary of Kabsch and Sander [6] which treats the problem in a stereochemical manner. Hydrogen bonds are defined on the basis of an energy criterion, then the β-sheet is constructed by identifying first bridges of two hydrogen bonds, followed by β-ladders of contiguous stretches of bridges and finally β-sheets of adjacent ladders. A ladder is composed of two β-strands hydrogen-bonded together. We make a distinction between *internal* and *external* β-residues on the basis of the number of β-ladders in which a residue participates. *Internal* residues belong to two ladders whilst *external* residues belong to a maximum of one.

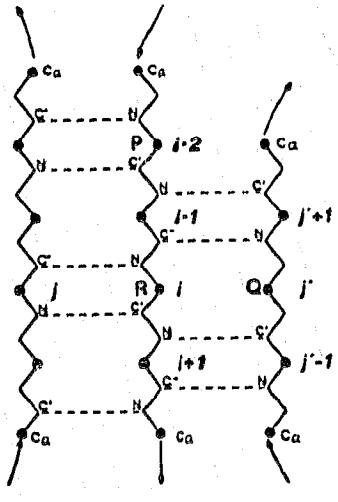

Fig. 1. Schematic representation of a three-stranded antiparallel β-sheet. Residue **R** at position *i* is *internal* according to the definition given in the text. Residues **P** and **Q** are *external*.

Table I

GOR prediction parameters for the *external* β-substate

	External β-substate																
	j-8		*j*-6		*j*-4		*j*-2		*j*		*j*+2		*j*+4		*j*+6		*j*+8
G	18	11	16	25	32	24	21	3	−21	−7	37	33	18	13	13	18	−19
A	−6	5	−6	−4	−11	−7	2	−20	−32	−22	−11	1	6	7	3	−2	10
V	−8	−17	−19	1	−3	6	29	62	57	37	28	9	−4	−14	1	9	−7
L	−22	−30	−40	−39	−51	−43	−24	−18	9	−5	−18	−11	−23	−11	−1	−1	−10
I	8	−19	−23	−12	−23	0	26	48	59	50	32	7	6	0	−44	−44	−42
S	19	25	28	30	41	38	20	21	17	15	15	17	15	15	12	12	32
T	28	37	9	13	13	28	23	22	33	34	18	17	36	38	11	2	15
D	−25	−16	8	24	15	14	−10	−50	−79	−76	−62	−17	−13	−21	−2	−31	−3
E	−19	−2	0	−10	−2	−27	−72	−53	−30	−20	−42	−38	−14	−4	4	−4	−13
N	20	18	33	17	9	11	−8	0	−48	−68	−38	−8	2	6	15	−4	4
Q	15	8	3	−17	−5	−47	−32	−29	13	10	−5	0	5	15	11	18	27
K	−32	−22	−31	−51	−22	−23	−43	−49	−26	−9	−4	4	−6	−11	−8	−5	0
H	−22	−13	−30	0	0	8	−4	−16	−26	−4	−8	8	11	−8	−4	12	19
R	−13	−10	−36	−16	−62	−57	−57	−103	−50	9	−9	−16	−28	−28	−6	−7	−37
F	7	−31	2	−10	21	21	16	24	24	18	14	−40	−47	−34	−38	−22	−19
Y	13	26	26	17	3	−25	12	23	19	14	21	−14	8	5	15	28	31
W	−15	7	−21	−14	−38	−38	40	58	18	12	−7	0	−79	−79	−79	−23	−50
C	−8	14	23	−23	−15	−7	10	10	28	17	−36	−38	−20	−24	14	14	4
M	−114	−113	−85	−63	−35	−62	−35	−6	0	119	−59	−119	−91	−91	−85	−35	−73
P	18	2	23	4	−13	4	−11	−46	−61	−52	−22	12	19	33	8	8	12

The information in centinats (nats $\times 10^{-2}$) that a residue at position *j* carries about the conformation of residues between positions *j*-8 and *j*+8 are given in columns *j*−8 through *j*+8, respectively. For example, all leucine residues carry −24 centinats of information about the amino acid two residues prior to the leucine in the sequence being in the *external* β conformation. Formally speaking therefore, the values quoted are those of $I(S_{j+m} = X : X; R_j)$, where $-8 \leq m \geq 8$, as originally used by Robson and Suzuki [9] and Garnier et al. [5] and not $I(S_j = X : X; R_{j+m})$ as quoted by Gibrat et al. [3]. The statistical measure of 'information' (the nat) is given by the natural logarithm of the ratio of the probability of a particular conformational state occurring at a certain position in the amino acid sequence given a particular amino acid at a second position in the sequence over the probability of this conformational state at the first position, independent of the type of amino acid in the second position. For small datasets a Bayes expected frequency method can be used (and has been adopted here) which employs a hash function in place of the logarithm as describe by Robson and Suzuki (Eqns 6, 7 and 8 of [9]).

Table II

GOR prediction parameters for the *internal* β-substate

	Internal β-substate																
	j-8		*j*-6		*j*-4		*j*-2		*j*		*j*+2		*j*+4		*j*+6		*j*+8
G	21	16	30	31	41	42	7	−44	−72	−34	−18	4	36	32	23	9	21
A	−27	−35	−14	−21	−7	−13	−9	0	−18	−31	−11	−23	−11	−4	−2	10	6
V	−28	−31	−16	140	−30	0	51	63	97	96	60	21	−18	−29	−37	−51	−18
L	−10	−13	−27	−39	−12	11	27	64	40	33	−2	−53	−32	−42	−43	−21	−21
I	−33	−28	−19	−23	3	25	46	70	83	76	42	7	−14	−19	−28	−47	−24
S	11	17	31	33	15	26	18	−27	−36	−17	4	14	41	37	45	37	22
T	43	23	45	28	13	5	20	18	13	16	25	20	−12	−28	8	3	0
D	27	34	14	6	31	−12	−42	−61	−120	−123	−37	8	23	21	14	21	8
E	−16	0	−11	−19	−47	−58	−42	−65	−78	−85	−53	−23	−37	−28	−24	−11	−4
N	−7	7	−15	3	3	−15	−32	−78	−160	−133	−74	−24	7	26	29	37	11
Q	0	27	0	18	−15	−5	−39	−46	−54	−46	−10	−10	−27	−27	0	14	27
K	−8	−30	−30	−8	−40	−43	−55	−74	−55	−74	−51	−8	8	13	20	3	10
H	0	−34	7	26	35	13	13	−15	−32	−63	−32	−15	−32	26	−25	−8	−25
R	21	6	11	−24	−12	−18	−6	−6	−18	−47	5	20	21	0	−12	−26	−26
F	−84	−44	−51	−9	−37	0	21	49	55	54	8	−9	−51	−75	−75	−51	−54
Y	9	22	18	22	9	17	13	42	70	62	45	57	25	29	−5	−15	−15
W	33	33	12	12	42	12	−28	−28	66	66	33	−66	−66	−46	−28	−46	0
C	−91	−51	−97	−58	−82	−29	−29	38	38	12	−20	−48	−135	−135	−58	29	13
M	−57	−12	−40	−25	−78	−12	−25	21	46	30	46	60	−12	−57	−104	−104	−12
P	18	34	4	8	14	−58	−104	−299	−299	−78	−12	0	27	38	35	14	−8

The values quoted are analogous to those given in Table I.

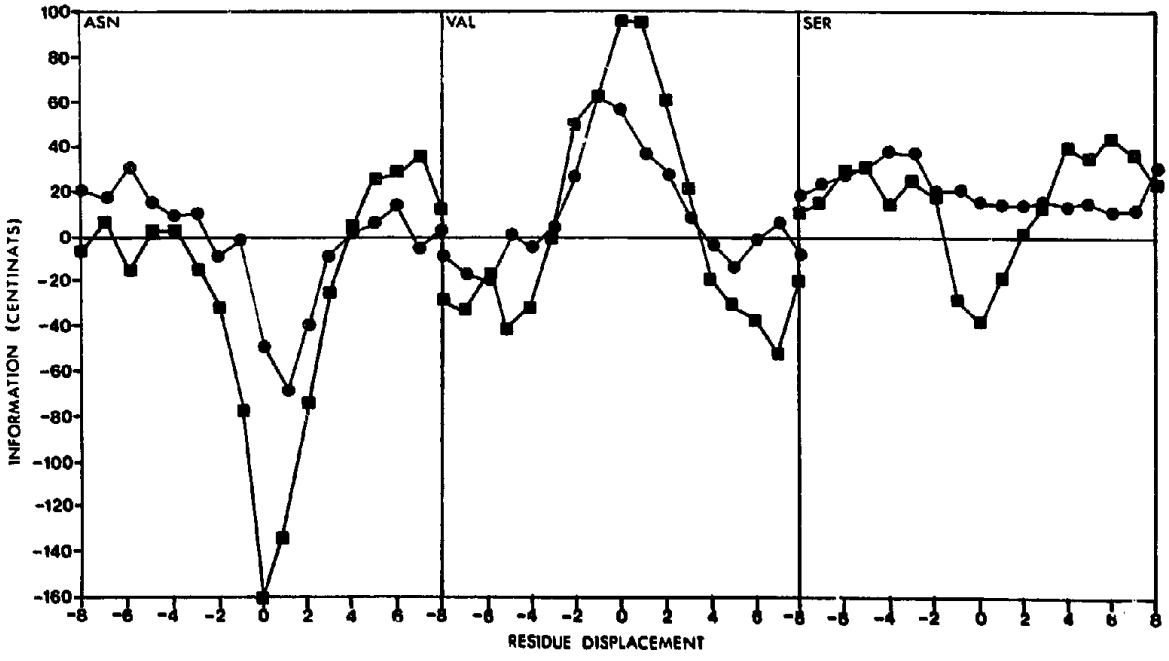

Fig. 2. Information profiles for (a) asparagine, (b) valine and (c) serine. Values are taken directly from Tables I and II. *Internal*-substate parameters (■) and *external*-substate parameters (•).

Using the nomenclature of Kabsch and Sander [6] an *internal* residue at position i in an antiparallel β-sheet will fulfill both of the following criteria:
(i) antiparallel bridge(i,j) = [H-bond(i,j) and H-bond(j,i)]
(ii) antiparallel bridge(i,j') = [H-bond($i-1,j'+1$) and H-bond($j'-1,i+1$)]

It thus forms a bridge composed of two hydrogen bonds with residue j on one strand and a second bridge with residue j' on a second strand (e.g. residue **R** in Fig. 1). The case for an internal residue (i) in the parallel arrangement is as follows:
(i) parallel bridge(i,j) = [H-bond($j-1,i$) and H-bond($i,j+1$)]
(ii) parallel bridge(i,j') = [H-bond($i-1,j'$) and H-bond($j',i+1$)].

All β-residues not contributing to two β-ladders are defined *external*.

Prediction parameters for use with the GOR directional method were derived for the β-substates from a dataset composed of 56 polypeptide chains. The Brookhaven codes [7] for the 53 coordinate datasets used were as follows: 2APP, 2ACT, 1ACX, 4ADH, 1ABP, 3CPV, 1CAC, 5CPA, 2CTS, 2CNA, 1CRN, 2B5C, 4CYT, 1CYP, 2CDV, 1ECD, 2FD1, 1FDX, 3FXC, 2GRS, 4LDH, 1NXB, 1SN3, 1PPT, 3PGK, 1BP2, 1LZM, 2LYZ, 1PCY, 1RHD, 1RN3, 3RXN, 2SNS, 1SBT, 3TLN, 2PTN, 4PTI, 3CAT, 2WGA, 4DFR, 1HMQ, 3FAB, 1REI, 1OVO, 2MDH, 1MLT, 2PAB, 2SOD, 1TIM, 1INS, 1PKA, 1SGB, and 1TGS. The dataset includes members of all of the four major structural classes of proteins, all-β, all-α, β + α and β/α [8]. The total number of residues in the dataset was 9874 including 1876 β-residues of which 642 are *internal* and 1234 *external*. Parameters were derived using the method of Robson and Suzuki [9] by application of the hash function to the observed frequency data extracted directly from the Kabsch and Sander definitions (Eqn 6 of [9]). Updated parameters for the prediction of the conventional states α-helix, β-sheet, turn and coil for use with the substate parameters were also derived from the same dataset by an analagous procedure to that described above. Tables of these parameters are not presented here but are similar to those reported recently by Gibrat et al. [3] and are available on request from the authors.

2.2. *Decision constants*

Initial predictions using zero decision constants [5] showed an over-prediction of the content of *internal* and *external* β-residues and α-helix (see Results). Optimal decision constants for these three states (DC_{ext}, DC_{int} and DC_{α}) were therefore calculated in order to give the correct percentage of each type of secondary structure when predictions were performed on the same dataset from which the parameters had been derived. Values were calculated by determining the percentage of each state predicted at all points in a 3-dimensional array in which DC_{ext}, DC_{int} and DC_{α} varied from −300 to +300 centinats in increments of 5 centinats. The coordinates of the single point (or where necessary interpolation between points) in the array which gave the correct percentage of each secondary structure were taken as the optimal decision constants.

3. RESULTS AND DISCUSSION

Tables I and II give the parameters for the two new β-substates. The benefit in making the distinction can be seen in Fig. 2 where the information profiles for asparagine, valine and serine are shown by way of example. The shapes of the profiles for *internal* and *external* residues are often similar but the magnitude of the major peaks can vary considerably. This is the case for both asparagine and valine. In the former case a strong preference against *internal* residues is noticeably less severe for *external* residues and the reverse is true of valine. In general the hydrophobic nature of the sheet core is borne out by the parameters with the hydrophobic amino acids dominating the *internal* positions. Others of the profiles including serine show a more complex behaviour. In this case the profile changes from a symmetric distribution, negative at its centre for *internal* residues to an entirely positive profile for *external* residues. This increase in information for external positions is expected on the basis of the

Table III

The predicted and observed abundance of secondary structural states

	Percent predicted	Percent observed	Optimised decision constant in centinats
α-helix	30.3	25.9	53
internal β-sheet	13.1	6.5	108
external β-sheet	15.3	12.5	55

Predictions were made using zero decision constants. Optimal decision constants to rectify the overprediction are also quoted.

hydrophilic nature of the serine side-chain. These distinctions in the profiles had been previously overlooked by considering all β-residues homogeneous.

3.1. PREDICTION

The parameters for the novel substates together with those for α-helix, turn and coil enable a 5-state prediction to be made. Here, however, we are primarily interested in the identification of regions of regular secondary structure and have considered only a four-state prediction in which coil and turn are taken to be a single state. Predictions were made on the same dataset from which the parameters had been derived and the percentage of each secondary structural state predicted using zero decision constants compared with their observed values are given in Table III. Also given are the decision constants necessary to eliminate the observed overprediction of each state.

Predictions using zero decision constants (i.e. those most commonly used and quoted in the literature) resulted in the correct prediction of 39% of the *external* residues and 58% of the *internal* residues when tested on the 53 protein trial dataset. The value for *internal* residues is comparable to that obtained for other states (we obtained for example 57% accuracy over all states) and yet provides more tertiary structural information. Use of the optimal decision constants improved the overall percentage correct from 57% to 61%. This compares favourably with 64% for a *three-state* prediction just using the updated parameters for α-helix, β-sheet and coil with their optimised decision constants. The advantage of the new parameters does not reside in an improvement of the overall quality of prediction but in the information provided by distinguishing the substates.

On use of the optimal decision constants to eliminate overprediction the percentage of β-sheet residues correctly predicted by the updated conventional parameters was 50% in comparison with 49% correctly predicted by the β-substrate parameters (the exact substate, *external* or *internal*, predicted being ignored for comparison with the known structure). 37% and 34% of *internal* and *external* residues, respectively, were correctly predicted by the β-substate parameters.

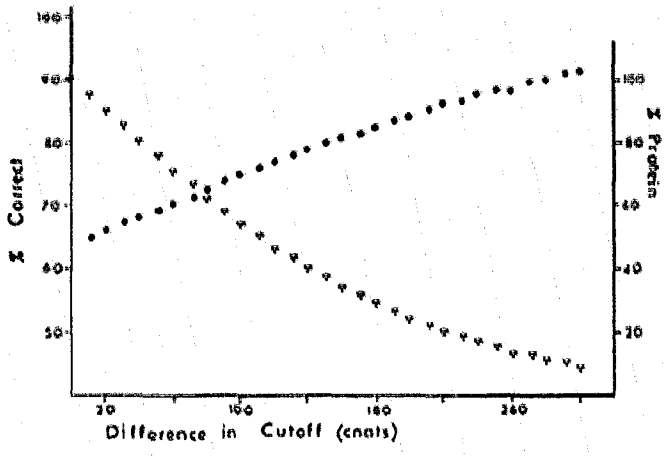

Fig. 3. Reliability of prediction. The percentage of residues above the cut-off (a) and the percentage of these residues correctly predicted (•) is plotted as a function of 'difference in information' cut-off. Predictions were made on the same dataset as that used for deriving the parameters.

Although 37% for the *internal* residues is an impressive figure for a state whose overall abundance is only 6.5% (Table III), it may impose a limitation on the application of such parameters. However, such figures can be misleading in that often a prediction is broadly correct but fails to locate accurately the termini of the elements of secondary structure. This may lead to a poor estimate of the quality of prediction on a residue for residue basis and yet for the purposes of the investigator may be a sufficiently accurate representation of the sequential topology of the molecule.

One means to overcome this problem is to enhance the reliability of at least part of the prediction by reference to the difference in the magnitude of the information value obtained for the predicted state over its next nearest rival at each position in the sequence. Fig. 3 shows how the percentage of all residues correctly predicted increased as a function of a 'difference in information' cut-off. At each point on the graph only those residues for which this difference in information value exceeded the cut-off were included in the analysis. Thus, for example, if an information cut-off of 150 centinats is applied an accuracy of almost 80% can be acheived but at the expense of only predicting the structure of ~40% of the molecule. A similar observation has been reported by Gibrat et al. [3].

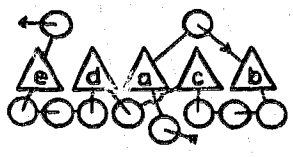

Fig. 4. Schematic topology diagram of adenylate kinase. Triangles represent β-strands and circles α-helices.

3.2. *An application β/α proteins*

To demonstrate the potential usefulness of the substate prediction in tertiary folding we have considered the β-sheet topologies observed in many of the β/α proteins. Typically the β-sheets of such proteins are all parallel and consist of 5 or 6 strands in which the first (sequential) strand is internal in the sheet and the final (sequential) strand is an edge strand. A topology diagram [10] for adenylate kinase [11] serves as an example and is given in Fig. 4.

If we make no *a priori* assumptions about a five stranded β-sheet other than that it is all parallel then there are 60 (5!/2) possible topologies which the sheet can adopt. This is based simply on the number of permutations of the five strands excluding the two-fold symmetry related topologies. However, if a prediction were able to locate the two edge strands unequivocally the problem of finding the correct topology would be reduced by an order of magnitude, namely to 6 (3!) possibilities, dramatically reducing the problem of predicting the tertiary fold. This type of problem is therefore particularly amenable to the application of the substate parameters.

Fig. 5 shows the results of the prediction of adenylate kinase using parameters derived solely from the following subset of β/α proteins: 4ADH, 2ADK, 2TAA, 1ATC, 3GPD, 3PGM, 4FXN, 8PAP, 4ADH, 1ABP, 1CAC, 5CPA, 2CTS, 2B5C, 2GRS, 3PGK, 1RHD, 1SBT, 3CAT, 4DFR, and 1TIM. The predictions again utilise optimised decision constants but employ no information cut-off. Two predictions are given, one incorporating the β-substates and one conventional prediction. Both fare equally well in identifying the positions of the β-strands in the sequence. The substate prediction suggests A and D to be internal and E is the strongest candidate for one of the edge positions. We know that the second edge strand is almost always the final one in the sequence and therefore can be confident that the assignment of E to this position is correct. Of the remaining strands (B and C) the prediction at C is not unreasonable for an internal strand because although two residues are wrongly assigned *external* they are at the terminus of the strand where they are commonly observed (see, for example, strand A). This leaves B as the remaining edge strand which, although this is only weakly predicted in the substate prediction,

```
Sequence... MEEKLKKSKIIFVVGGPGSGKGTQCEKIVQKYGYTHLSTGDLLRAEVSSGSARGKMLSEIMEKGQLVPLETVLDMLRDAM
Strands....              A                       B
Structure..  HHHHH   XYYYXX     HHHHHHHHHHHH   XXXXHHHHHHHHHHHH   HHHHHHHHHH        HHHHHHHHHHHH
Pred. 1....  HHHHHH  EEEEEEE        HHHHHHHE   EE    H HHEH     HHHHHHHHHH        HHHHHHHHHHHH
Pred. 2....  HHHHHH  XXYYYX         HHHHHHX   X     H HHHH     HHHHHHHHHH        HHHHHHHHHHHH

Sequence... VAKVDTSKGFLIDGYPREVKQGEEFERKIGQPTLLLYVDAGPETMTKRLLKRGETSGRVDDNEETIKKRLETYYKATEPV
Strands....              C                       D
Structure.. H        YYYY     HHHHHHHHHHH     XYYYYX  HHHHHHHHHHHHHHHH        HHHHHHHHHH HHH
Pred. 1.... HHE     EEEE       HHH HHHHH     HEEEE     HHHHHHHH     E     HHHHHHHHHHH    HE
Pred. 2.... HHX     XXYY       HHH HHHHH     HYYYY     HHHHHHHH           HHHHHHHHHHH    HX

Sequence... IAFYEKRGIVRKVMAEGSVDDVFSQVCTHLDTLK
Strands....              E
Structure.. HHHH     XXXXX    HHHHHHHHHHHHHH
Pred. 1.... HHHH     EEEE         EE      HHH
Pred. 2.... HXHH     XXXX         Y       HHH
```

Fig. 5. The prediction of the secondary structure of adenylate kinase. The position of the five β-strands, A–E are shown boxed. The row labelled 'structure' shows the Kabsch and Sander assignments [6] of secondary structure simplified to show only α-helix, H; *external* β-sheet; X; and *internal* β-sheet, Y. All other positions are left blank. Prediction 1 was made using only updated conventional parameters and prediction 2 included the use of the β-substate parameters.

is corroborated by the conventional prediction. The problem would now be reduced to that of predicting the correct topology for the three *internal* strands.

4. CONCLUSION

We have introduced a simple modification to the GOR method for secondary structure prediction of proteins in an attempt to introduce some rudimentary tertiary structural information namely the identification of *internal* β-strands (core strands) as distinct from *external* β-strands (edge strands). We have shown the potential advantages of such information in the restriction of possible sheet topologies by reference to the β/α proteins. Similar arguments could be made for other structural classes. For example, it may be possible to identify Greek key motifs [12] in all-β proteins since these have a defined sequential arrangement of *internal* and *external* strands.

The most powerful adaptation of the method would be the use of a template fitting procedure for identifying $\beta_{int}\alpha\beta_{ext}$ or $\beta_{ext}\alpha\beta_{int}$ motifs thereby locating edge strands which are sequentially internal. The framework for such a modification has been provided by the algorithm of Taylor and Thornton [13,14].

REFERENCES

[1] Busetta, B. and Hospital, M. (1982) Biochem. Biophys. Acta 701, 111–118.
[2] Kabsch, W. and Sander, C. (1983) FEBS Lett. 155, 179–182.
[3] Gibrat, J.-F., Garnier, J. and Robson, B. (1987) J. Mol. Biol. 198, 425–443.
[4] Garratt, R.C., Taylor, W.R. and Thornton, J.M. (1985) FEBS Lett. 188, 59–62.
[5] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[6] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[7] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 122, 535–542.
[8] Levitt, M. and Chothia, C. (1976) Nature 261, 552–557.
[9] Robson, B. and Suzuki, E. (1976) J. Mol. Biol. 107, 327–356.
[10] Sternberg, M.J.E. and Thornton, J.M. (1977) J. Mol. Biol. 155, 1–17.
[11] Dreusicke, D., Karplus, P.A. and Schulz, G.E. (1988) J. Mol. Biol. 199, 359–371.
[12] Richardson, J.S. (1977) Nature 268, 495–500.
[13] Taylor, W.R. and Thornton, J.M. (1983) Nature 301, 540–542.
[14] Taylor, W.R. and Thornton, J.M. (1984) J. Mol. Biol. 173, 487–514.