The influence of tertiary structure on secondary structure prediction

Accessibility versus predictability for β-structure

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

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

Received 3 June 1985

The secondary structure prediction algorithm of Garnier et al. [(1978) J. Mol. Biol. 120, 97–120] has been used for 16 proteins whose structures are dominated by β-sheet. Comparisons of the predicted structures with those defined by the algorithm of Kabch and Sander [(1983) Biopolymers 22, 2577–2637] shows that for β-sheet residues, the quality of prediction falls markedly with increasing residue accessibility. 2 sub-classes of β-residues have been distinguished on the basis of hydrogen bonding patterns, and the distribution of amino acid types within each sub-class found to be quite different. Accordingly, Chou and Fasman P_β-type parameters for these previously indistinguished states have been derived.

Secondary structure prediction *β-Sheet residue* *Tertiary structure* *Solvent accessible area* *Residue hydrogen bonding*

1. INTRODUCTION

The success of protein secondary structure prediction by simple statistical algorithms currently stands at under 60% for a 3-state prediction (α-helix, β-sheet and coil) [1,2]. The failure of such methods [3,4] to improve is most often attributed to the constraints imposed by tertiary structure and long-range interactions [5]. For example, the 'hydrophobic' effect is a powerful constraint, leading to the formation of hydrophobic cores and more hydrophilic exteriors. The observed secondary structure is a balance between the conformational preferences of individual amino acids and the requirement to form a compact globular structure.

To approach this problem we have studied a sub-group of proteins, the structures containing predominantly β-sheet, to see if there is any correlation between tertiary structure and the failure of secondary structure prediction. Initial visual inspection of these structures showed that it is predominantly the edge strands of β-sheets and the end residues of β-strands, which are not located. To analyse these data in more detail we have searched for a correlation between the prediction accuracy for β-structure and the observed residue solvent accessibility and hydrogen-bonding. The data suggest that it is useful to segregate β-residues into 2 sub-classes (internal and external). We show that there are large differences, between these 2 classes, in the β-forming potential (P_β) [6] for certain amino acids.

2. METHODS

The secondary conformations of 16 proteins of known crystal structure were predicted using the directional method of Garnier et al. [4] with no decision constant bias. Predicted structures were

* To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

compared against those defined from all atom coordinates by the algorithm of Kabch and Sander [7]. Whilst predictions were made on a 4-state basis (α-helix, β-sheet, coil and reverse turn), for the purpose of comparison this was reduced to only 3 states, α-helix, β-sheet, and 'coil' (where coil is here taken to mean 'non-helix, non-sheet'). In all cases the quality index used was that of the percentage of residues whose structure is correctly predicted. The following proteins were selected as being representative of those whose secondary structure is dominated by β-sheet; Bence-Jones immunoglobulin REI variable portion (REI), *Streptomyces* subtilisin inhibitor (SSI), plastocyanin (PCY), penicillopepsin (APP), γ-crystallin II (CRS), ribonuclease S (RNS), azurin (AZU), human plasma prealbumin (PAB), γ-chymotrypsin A (GCH), tosyl elastase (EST), neurotoxin (NXB), actinidin (sulphydryl protease) (ACT), concanavalin A (CNA), λ-immunoglobulin Fab′ light chain (FAB), Cu,Zn superoxide dismutase (SOD), and papain (PAP). Altogether they contain a total of 2755 residues, including 272 β-sheet strands. Initially, topology diagrams [8] were used to identify regions where the prediction was failing badly, leading to 2 further approaches, each to investigate the quality of prediction in the context of residue position within the tertiary structure.

2.1. *Solvent accessible areas*

The program of Kabch and Sander [7] returns the static solvent accessible area (in $Å^2$) for each residue within a protein. We have computed for each residue in the dataset the dimensionless quantity of 'relative accessibility', i.e. the ratio of the residue accessible area in the protein to that in the X position of a theoretical tripeptide Gly-X-Gly. For this purpose the values of maximal accessible area for the central residue of the tripeptide given by Chothia [9] have been used. Although Chothia [9] used the rolling ball approach of Richards [10], and Kabch and Sander use 'geodesic sphere integration', they both assume identical Van der Waals radii and the 2 approaches are known to give good agreement when identical parameters are used [7,11]. Since terminal residues can theoretically have accessible areas greater than the central residue of the tripeptide, these were disregarded. Relative accessibilities were considered in ranges of 10% (0–10, 10–20%, etc.) and the percentage of residues correctly predicted computed in each range. The procedure was performed firstly over all residues and secondly over β-sheet residues alone.

2.2. *'Internal' and 'external' β-residues*

The program of Kabch and Sander [7] assigns β-secondary structure in a hierarchical manner, initially hydrogen bonds, then bridges, ladders and sheets. A ladder is formed by a pair of strands hydrogen-bonded together. We identify 2 classes of β-strand residue, termed 'internal' and 'external' β-residues on the basis of the number of β-ladders in which a given residue participates. Internal β-residues are defined as belonging to 2 ladders, whilst external β-residues belong to a maximum of 1. Residue R (at position i) (fig.1) is an example of an internal residue for the simple double antiparallel case. Using the Sander nomenclature [7], it fulfils both the following criteria:

(a) Antiparallel bridge (i,j) =
[H-bond (i,j) and H-bond (j,i)]

(b) Antiparallel bridge (i,j') =
[H-bond $(i-1,j'+1)$ and
H-bond $(j'-1,i+1)$]

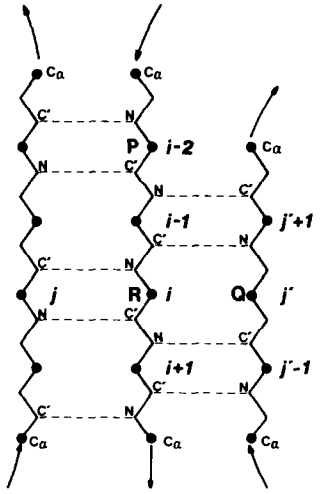

Fig.1. A hypothetical section of β-sheet showing 2 antiparallel ladders. Residue R (at position i) is classed 'internal', and residues P and Q (at $i-2$ and j', respectively) are classed 'external' according to the definition in the text.

In contrast, residue P (at $i-2$) fulfils only (a) and residue Q (at j') fulfils only (b) and both are therefore examples of external β-residues. All β-residues in strands at the edge of β-sheets will by definition be external, and similarly residues near the terminus of any strand are most likely to be external because of the general inequality in length of strands within a sheet, hence the terms external and internal.

We have computed the quality of prediction for both classes of β-residue, in each of the 16 proteins used. In addition, the distribution of the 20 amino acids in the 2 classes has been determined and used to derive Chou and Fasman propensity parameters [6] for each. We term such parameters $P_{\beta i}$ for internal and $P_{\beta e}$ for external β-residues.

3. RESULTS AND DISCUSSION

Examples of the simplified topology diagrams are shown in fig.2 for prealbumin and immunoglobulin Fab' light chain variable domain, where major errors in prediction are shown hatched. Consideration of these indicated that predictions failed badly most often in edge sheet strands. Such strands are often short and hydrophilic [12], and will tend to occur at the molecular surface where their solvent exposure would be expected to be great. This is dramatically demonstrated in fig.3 where the quality of prediction for β-residues is seen to fall markedly with increasing relative accessibility, the downward trend being significant at the 0.01 level (see figure legend). In contrast when all residues, irrespective of defined structure, are considered, no such trend is observed.

Of the 333 internal and 623 external β-residues in the dataset, 68.2 and 44.1% respectively, were correctly predicted. Paired t-statistics show the mean difference of 23.5% across the 16 proteins to be significant at the 0.01 level. The evaluated Chou and Fasman propensity parameters for the 2 subclasses are shown in fig.4 where superimposed histograms for $P_{\beta i}$ (solid) and $P_{\beta e}$ (hatched) are plotted for the 20 amino acids in order of descending $P_{\beta i}$. For clarity only the difference between the parameters is indicated for each amino acid, the shading being that for the higher value of

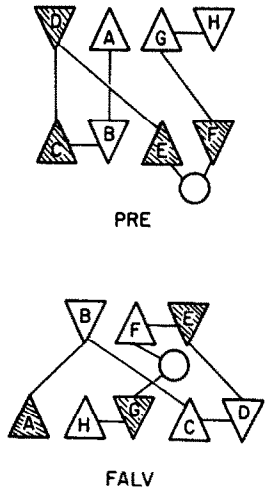

Fig.2. Schematic diagrams for (a) prealbumin, (b) immunoglobulin light chain variable domain from Fab' (new); β-strand regions which are incorrectly predicted using the standard Robson method [4] are indicated by hatching. For these triangle and circle diagrams each β-sheet is viewed along the strand direction. Each β-strand is represented by a triangle whose apex points up or down according to whether the strand is viewed from the N or the C end.

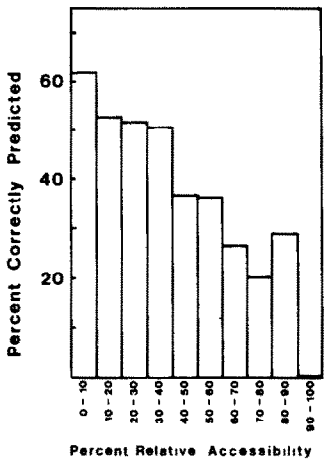

Fig.3. Histogram showing the success of prediction for β-residues against residue 'relative accessibility' (defined as the ratio of the static solvent accessible area of the residue in the protein to that in the X position of a hypothetical tripeptide Gly-X-Gly). The downward trend tested against a null hypothesis of invariant success with change in 'relative accessibility', using the F statistic with $2N-2$ degrees of freedom ($N = 10$, the number of relative accessibility ranges) is significant at the $P < 0.01$ level.

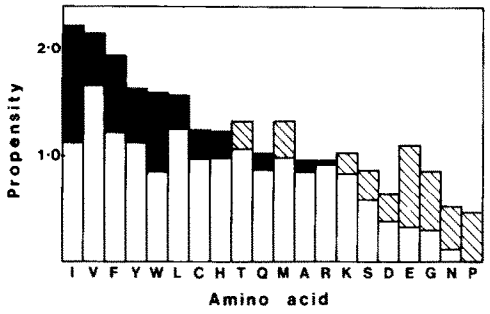

Fig.4. Superimposed histograms of $P_{\beta i}$ (solid) and $P_{\beta e}$ (hatched) for the 20 amino acids, in order of descending $P_{\beta i}$. For clarity only the difference between the parameters is indicated for each amino acid, the shading being that of the higher value of P. Variance ratio analysis shows $s^2P_{\beta i}$ (0.426) and $s^2P_{\beta e}$ (0.080) to be significantly different at the $P < 0.001$ level.

P. The clustering of solid blocks to the left of the diagram and hatched blocks to the right is striking, indicating that the strong β-forming amino acids show a preference for internal positions, where the converse is true of weak β-forming amino acids. The variance in $P_{\beta i}$ is significantly greater than that in $P_{\beta e}$ at the 0.001 level. In the case of certain amino acids (notably Ile, Gly and Glu) the $P_{\beta i}$ and $P_{\beta e}$ propensities differ by a factor of approx. 2, involving absolute changes in P of greater than 1. As anticipated, proline is never found in an internal position.

The 2 independent criteria to classify β-residues (i.e., solvent accessibility and H-bonding patterns) give equivalent results, suggesting that there is a different sequence requirement for internal and external strands. The prediction algorithm into which such information can most readily be incorporated is that of Garnier et al. [4] rather than that of Chou and Fasman [3] and we are currently deriving the appropriate parameters for β_i and β_e which will discriminate between these internal and external residues and hopefully improve the prediction. Potentially, such an assignment could also be used to restrict the possible topologies in tertiary structure prediction.

ACKNOWLEDGEMENTS

R.C.G. was the recipient of an SERC advanced course studentship and W.R.T. holds an SERC advanced fellowship.

REFERENCES

[1] Kabch, W. and Sander, C. (1983) FEBS Lett. 155, 179–182.
[2] Busetta, B. and Hospital, M. (1982) Biochim. Biophys. Acta 701, 111–118.
[3] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 222–245.
[4] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[5] Kabch, W. and Sander, C. (1984) Proc. Natl. Acad. Sci. USA 81, 1075–1078.
[6] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 212–222.
[7] Kabch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[8] Sternberg, M.J.E. and Thornton, J.M. (1976) J. Mol. Biol. 105, 367–382.
[9] Chothia, C. (1976) J. Mol. Biol. 105, 1–14.
[10] Richards, F.M. (1977) Annu. Rev. Biophys. Bioeng. 6, 151–176.
[11] Lee, B.K. and Richards, F.M. (1971) J. Mol. Biol. 55, 379–400.
[12] Sternberg, M.J.E. and Thornton, J.M. (1977) J. Mol. Biol. 155, 1–17.