J. Mol. Biol. (1993) **234**, 951–957

The Limits of Protein Secondary Structure Prediction Accuracy from Multiple Sequence Alignment

Robert B. Russell and Geoffrey J. Barton†

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

(Received 15 April 1993; accepted 16 August 1993)

The expected best residue-by-residue accuracies for secondary structure prediction from multiple protein sequence alignment have been determined by an analysis of known protein structural families. The results show substantial variation is possible among homologous protein structures, and that 100% agreement is unlikely between a consensus prediction and one member of a protein structural family. The study provides the range of agreement to be expected between a perfect secondary structure prediction from a multiple alignment and each protein within the alignment. The results of this study overcome the difficulties inherent in the use of residue-by-residue accuracy for assessing the quality of consensus secondary structure predictions. The accuracies of recent consensus predictions for the annexins, SH2 domains and SH3 domains fall within the expected range for a perfect prediction.

Keywords: secondary structure; prediction; sequence alignment

There are now a large number of proteins which share similar sequence, 3D‡ structure and function. Frequently, one or more of the members have a known 3D structure, making approximate structures of the other family members available by homology modelling (e.g. Blundell *et al.*, 1987). However, when 3D structural information, whether from X-ray crystallography, NMR or other experimental techniques, is not available for any members of a given protein family, 3D structural information must come from analysis of sequence alone. Accurate prediction of the protein secondary structure provides a valuable guide for experimental design when structure determination is difficult, or years from completion. In addition to providing an accurate starting point for tertiary structure prediction, such predictions may suggest which site-directed mutations are likely to disrupt the native fold (e.g. Russell & Barton, 1992), or identify the surface peptides most likely to be antigenic (e.g. Sternberg, *et al.*, 1987).

Recently, the traditionally poor performance of secondary structure prediction ($\approx 63\%$ accuracy (three-state; α-helix, β-strand, coil) on average (Holley & Karplus, 1989)) has been improved by the use of aligned protein sequence families (Rost & Sander, 1992; Barton & Russell, 1993; Thornton *et al.*, 1991; Russell *et al.*, 1992; Barton *et al.*, 1991; Crawford *et al.*, 1987; Rost & Sander, 1993; Rost *et al.*, 1993; Benner & Gerloff, 1991; Bazan, 1990; Zvelebil *et al.*, 1987). This has given improvements both in percentage accuracies, and the prediction of the number, type and location of secondary structures. However, since it is unusual for the experimentally determined secondary structure to be identical in all members of a protein family, a consensus prediction will rarely attain an accuracy of 100% for all family members. Here we use the secondary structure variation observed within protein structural families to determine the limits of residue-by-residue accuracy for secondary structure prediction from multiple alignment. We provide a protocol for estimating the range in expected accuracy for a perfect prediction given the sequence variation within the family. The protocol provides an improved means of assessing prediction accuracy, and shows that the accuracies of many recent predictions are within the expected range. The analysis also confirms that there can be substantial variation in secondary structure between homologous proteins.

Techniques of secondary structure prediction from multiple sequence alignment vary, but the common theme is the prediction of a consensus, or

† To whom correspondence should be addressed.

‡ Abbreviations used: 3D, three-dimensional; NMR, nuclear magnetic resonance; Ig, immunoglobulin.

0022 2836/93/240951–07 $08.00/0

core set of secondary structures for the entire family. For a single protein, the residue-by-residue accuracy of a secondary structure prediction is normally expressed as the percentage of correctly assigned residues, where the best possible result is 100%. However, within a family of protein structures, secondary structure variation is expected. The ends of helices and strands will often differ across the family, and small elements of secondary structure may be present only in some of the family members. Thus when comparing even a perfect prediction of the family's core secondary structures to any one member of the family, the accuracy will rarely be 100%. To estimate best prediction accuracy given an alignment of a particular length and composition, we have obtained structurally derived alignments (Russell & Barton, 1992) for 14 protein families, and compared the assigned secondary structure (DSSP; Kabsch & Sander, 1983) variation to the observed variation in sequence conservation.

The improved accuracy of secondary structure predictions made using multiple sequence alignments stems from the presence of conserved positions that indicate α-helix or β-strand and the presence of insertions/deletions indicating loops (Zvelebil *et al.*, 1987). The success of these methods thus depends on alignments containing sequences of varied composition. Very similar sequences readily yield accurate alignments, but patterns of conservation may not be clear, since most positions will be conserved. Distantly related sequences can yield clearer patterns of conservation, but may be difficult to align accurately, which leads to errors in the prediction. Sequence alignments best suited to predicting secondary structure fall between these two extremes. Secondary structure agreement varies as a function of the degree of conservation: proteins with similar sequences show little variation in secondary structure, whereas distantly related proteins show substantial secondary structure variation outside of the conserved core. The degree of conservation thus provides a means to measure both the expected predictive usefulness of the alignment and a scale on which to plot the expected accuracy of secondary structure prediction. We define conservation, C, as the percentage of alignment positions sharing seven or more property states (hydrophobic, aliphatic, not-charged, etc.) as defined by Zvelebil (Zvelebil *et al.*, 1987; Livingstone & Barton, 1993) across all aligned sequences.

Multiple protein sequence alignments vary in sequence composition, alignment length and in the number of sequences that they contain. Variation due to the number of sequences was removed by considering alignments of five sequences, and the effect of alignment length on both amino acid and secondary structure conservation was accounted for by defining four length ranges ($\leqslant 50$; 51 to 100; 101 to 150; and > 150).

Figure 1 shows how maximum and minimum consensus secondary structures may be obtained from a sequence alignment derived by 3D structure comparison. The two types of consensus provide a range over which a perfect secondary structure prediction is likely to fall. The average agreement of each secondary structure within the alignment with the maximum and minimum consensus provides an estimate of the best accuracy for a prediction made from the alignment. Figure 1(b) illustrates one method by which a prediction of secondary structure might be made from a multiple sequence alignment (Russell *et al.*, 1992).

The relationship between secondary structure agreement to perfect (alignment derived) prediction, and C is shown in Figure 2. Each point corresponds to the average agreement between one protein in the family and the maximum and minimum consensus defined in Figure 1(a). The accuracy of a perfect prediction is rarely better than 95%, with the lower range in accuracy increasing with increasing C. Four alignment length ranges were defined since the expected range in accuracy is a function of length: short alignments have a larger range than longer alignments. The figure provides a means of estimating the best possible success rate of the prediction from a sequence alignment.

The study confirms that a significant degree in secondary structure variation is found even among related protein structures (e.g. Lesk & Chothia, 1980). For example, when an alignment of six divergent globin sequences (Russell & Barton, 1992) is examined, a value between 23% and 28% is observed for C, and the observed agreement between each secondary structure and the minimum and maximum consensus is 79% to 88%. A prediction of the secondary structure for this family of sequences may be considered successful if it achieves an accuracy within this range.

We propose the following protocol to determine the expected accuracy of a perfect prediction made using a protein sequence alignment.

1. Select a sub-alignment containing the five most varied sequences among the family to be used in the prediction.
2. Calculate C according to Zvelebil *et al.* (1987).
3. Given the alignment length, refer to the appropriate plot within Figure 2 to determine the range of secondary structure variation expected, for C as determined in 2.

For example, for an alignment of length 120, with $C = 34\%$, Figure 2c gives an expected range of secondary structure consensus agreement between $\approx 80\%$ and 100% (100% is always the theoretical best). This means that the secondary structure of at least one protein from the alignment will show only 80% agreement with the consensus. The quality of a secondary structure prediction from this alignment should be judged accordingly.

The results of applying the above protocol to sequence families used in five recent predictions are shown in Table 1. For each family of N sequences, with alignment length L, and percentage conservation C, the Table shows how the obtained prediction accuracy compares with the best possible accuracy.

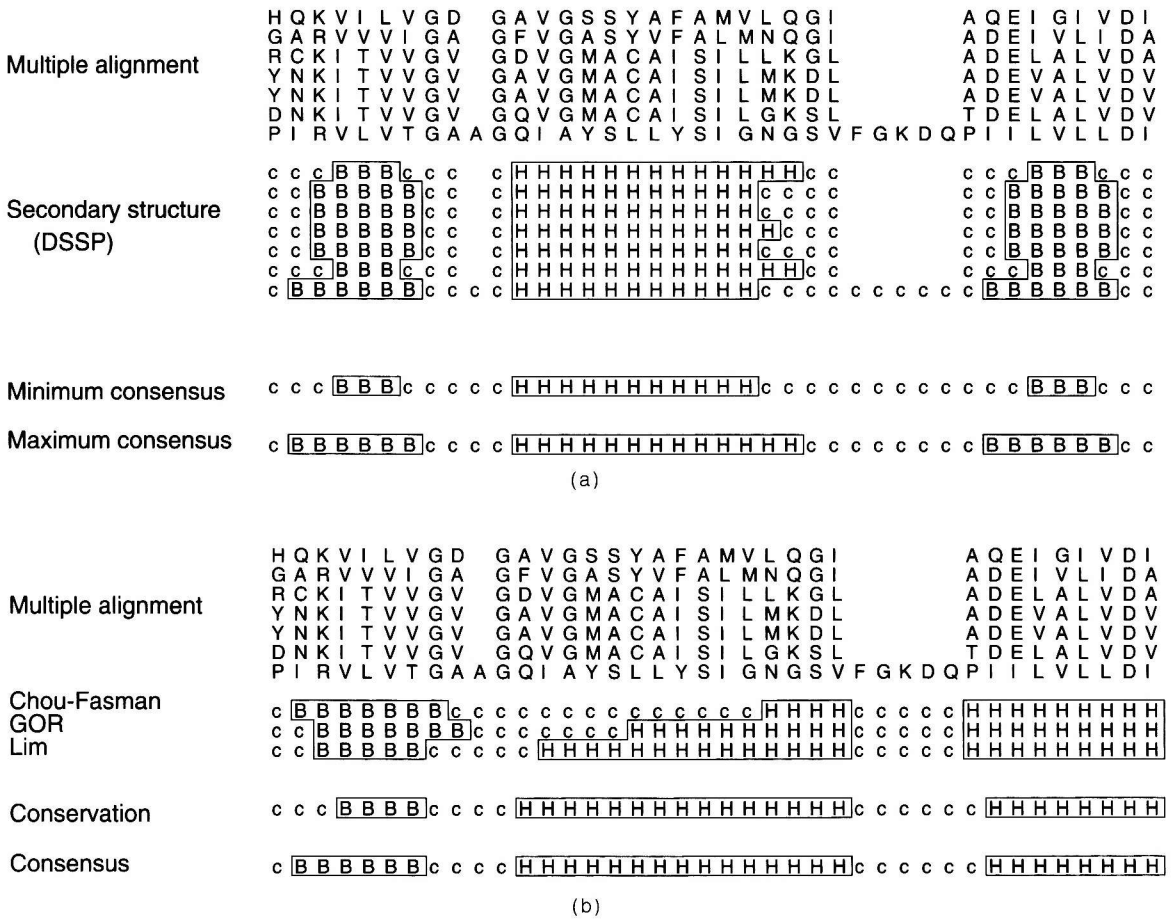

Figure 1. (a) The Figure shows (3-state) DSSP (Kabsch & Sander, 1983) secondary structure assignments, and how 2 types of consensus secondary structure assignments are defined from aligned proteins of known 3D structure. A maximum consensus shows which of helix (H) or beta (B) structure is present in any member of the family for each position in the alignment. Positions having both H and B, and positions having no H or B are labelled coil (c). A minimum consensus shows the positions where H and B are common across every member of the family, with all other positions labelled c. (b) An example of how a consensus secondary structure prediction might be derived. Three methods of secondary structure prediction (Garnier *et al.*, 1978; Lim, 1974; Chou & Fasman, 1978) are combined with a conservation pattern based prediction (Russell *et al.*, 1992), to give a consensus prediction, defined as a string of 3-state residue-by-residue predictions for each position within the alignment. In all predictions based on multiple alignment, residues can be defined as core secondary structures (helix, H or beta, B), or coil structure (c), providing a consensus similar to those defined in (a).

Of the 28 comparisons of predicted and experimental structures, 16 fall within the range of accuracy expected, suggesting that they are near perfect. Furthermore, the remaining predictions are more encouraging when judged beside the expected range of accuracy defined in Figure 2. For example, the apparently disappointing 56% residue-by-residue accuracy (Rost & Sander, 1992; Barton & Russell, 1993; Robson & Garnier, 1993) of the SH3 domain prediction of Benner (Benner *et al.*, 1992; 1993) should be viewed beside the possible minimum agreement of 70% for an alignment-based prediction of this family of proteins.

Secondary structure prediction from multiple protein sequence alignment predicts only the core secondary structures. When compared to an individual protein, such a prediction is incomplete. This study provides an appropriate measure by which to assess the success of prediction once experimentally determined structures are known for one or more of the proteins in the family. Variation in the lengths of secondary structures and structural content of loops can lead to a low residue-by-residue secondary structure prediction accuracy. Some authors have suggested assessing accuracy using secondary-structure element agreement (i.e. whether helix or sheet is predicted within the correct region) (Taylor & Thornton, 1983; Rost & Sander, 1992) since residue-by-residue accuracy can give apparently poor values even for good predictions. Although it is still desirable to determine whether a prediction has correctly predicted the number, type and location of secondary structure elements, the results of this study suggest that residue-by-residue accuracy can be an effective measure of the quality of an alignment based prediction.

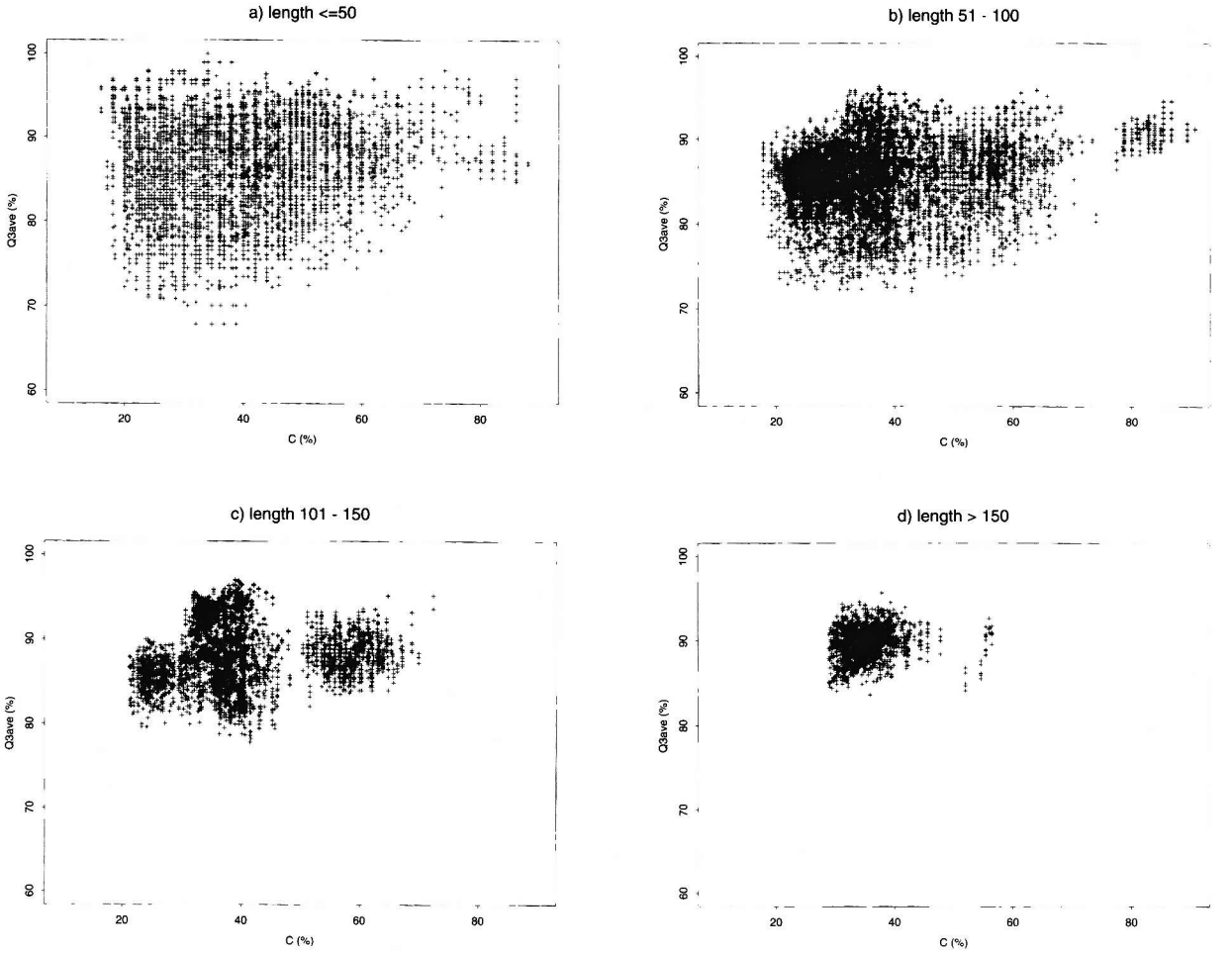

Figure 2. Plots of the average agreement between secondary structure assignments for each protein and the maximum and minimum consensus, (Q_{3ave}), *versus* percentage conservation (C) for alignments of 5 sequences taken from 14 protein structural families. Since both C and secondary structure agreement are dependent on length, the plots are divided into alignment length ranges: a, $\leqslant 50$ residues; b, 51 to 100 residues; c, 101 to 150 residues; and d, >150 residues. A single member from each structure family was used to scan (Russell & Barton, 1992) the current Brookhaven (Bernstein *et al.*, 1977) database (including pre-release) to find proteins related structurally. A representative structure (highest resolution, well-refined) was chosen for each structural sub-family having 90% sequence identity. Families were only considered if accurate alignment of the sequences without consideration of 3D structural information was possible. Unrefined structures and/or those of resolution greater than 2·5 Å were ignored. The viral coat proteins were included despite often having resolution greater than 2·5 Å since molecular averaging makes their structures of a similar quality to those of higher resolution. The structures used (Brookhaven codes in parentheses; chains are given after an underscore): (1) Ig heavy chain variable domains (1MAM H residues 1 to 123, 1IGM H residues 1 to 129, 8FAB_B residues 1 to 123, 1HIL B residues 1 to 115, 2FB4_H residues 1 to 120, 1FDL_H residues 1 to 118, 7FAB H residues 1 to 119, 2FBJ_H residues 1 to 122, 6FAB_H residues 301 to 423); (2) Ig heavy chain constant domains (7FAB_H residues 120 to 217, 8FAB_B residues 124 to 222, 6FAB_H residues 424 to 522, 1FDL_H residues 119 to 218, 1HIL B residues 116 to 228, 2FB4 H residues 121 to 218); (3) Ig light chain variable domains (7FAB_L residues 1 to 107, 2RHE all residues, 2FB4 L residues 1 to 113, 2MCG_1 residues 1 to 115, 8FAB_A residues 3 to 109, 1IMM residues 1 to 108, 1HIL A residues 1 to 111, 1IGM_L residues 1 to 115, 1FDL_L residues 1 to 111, 2FBJ_L residues 1 to 110, 6FAB_L residues 1 to 111); (4) Ig light chain constant domains (6FAB_L residues 112 to 214, 1FDL L residues 112 to 214, 2FBJ L residues 111 to 212, 1HIL_A residues 112 to 211, 2FB4_L residues 114 to 214, 7FAB_L residues 108 to 204, 2MCG_1 residues 116 to 216, 8FAB A residues 110 to 208); (5) Ig variable domains (families 1 & 3); (6) Ig constant domains (families 2 & 4); (7) globins (2LH1, 4MBN, 4HHB A, 4HHB_B, 1ECA, 1MBA, 2LHB, 1PMB A, 1FDH_G, 1PBX_A, 1PBX B, 1ITH_A, 1HBG, 2SDH A); (8) serine proteases (2PTN, 2PKA_AB, 1TON, 3RP2_A, 3EST, 4CHA_A, 1HNE_E, 1SGT); (9) aspartyl protease N terminal domains (3APP residues 1 to 174, 4APE residues −2 to 174, 2APR residues 1 to 178, 4PEP residues −2 to 174, 1CMS residues 1 to 175, 1RNE residues −1 to 172); (10) aspartic protease C terminal domains (3APP residues 175 to 323, 4APE residues 175 to 326, 2APR residues 179 to 325, 4PEP residues 175 to 326, 1CMS residues 176 to 323, 1RNE residues 176 to 323); (11) cytochrome *c* structures (1C2R_A, 1YCC, 5CYT_R, 1CCR, 1CYC); (12) viral coat proteins VP1 (2MEV_1, 1TME_1, 4RHV_1, 2PLV_1, 1R1A_1); (13) viral coat proteins VP2 (2MEV_2, 1TME 2, 4RHV 2, 2PLV_2, 1R1A_2); (14) viral coat proteins VP3 (2MEV 3, 1TME_3, 4RHV_3, 2PLV_3, 1R1A_3). Alignments were generated by using the STAMP package (Russell & Barton, 1992). Gaps between un-gapped segments of greater than 3 residues were adjusted to make their length minimal. A long insertion of 36 residues in the VP1 family (12) was shortened to 4 residues to prevent this gap from distorting the agreement of secondary structure assignment to the

Table 1

Recent predictions and their expected and observed accuracies

Sequence family	*N*	*L*	*C*	Expected accuracy (%)	Prediction	Structure(s) known	SS	Observed accuracy (%)
Trp synthase α	9	286	39·9	80–100	Crawford *et al.* (1987)	Hyde *et al.* (1988)	A	74
Kinase	89	417	20·4	80–100	Benner & Gerloff (1991)	Knighton *et al.* (1991)	A	63●
Annexin	88	90	28·9	70–100	Barton *et al.* (1991)	Huber *et al.* (1992)	A	79
						Bewley *et al.* (1993)	A	79
						Weng *et al.* (1993)	A	75
					Taylor & Geisow (1987)	Huber *et al.* (1992)	A	81
						Bewley *et al.* (1993)	A	79
						Weng *et al.* (1993)	A	84
SH2 domain	67	93	23·7	70–100	Russell *et al.* (1992)	Waksman *et al.* (1992)	A	78
						Eck *et al.* (1993)	A	80
						Overduin *et al.* (1992)	A	76
						Booker *et al.* (1992)	A	74
					Panayotou *et al.* (1992)	Waksman *et al.* (1992)	A	73
						Eck *et al.* (1993)	A	76
						Overduin *et al.* (1992)	A	75
						Booker *et al.* (1992)	A	78
SH3 domain	67	66	18·2	70–100	Benner *et al.* (1992)	Musacchio *et al.* (1992)	D	56●
					Rost & Sander (1992)	Musacchio *et al.* (1992)	D	70●
						Musacchio *et al.* (1992)	A	68
						Yu *et al.* (1992)	A	69
						Kohda *et al.* (1993)	A	57
						Koyama *et al.* (1993)●●	A	59
						Noble *et al.* (1993)	A	73
					Benner & Gerloff (1993)	Musacchio *et al.* (1992)	A	46
						Yu *et al.* (1992)	A	58
						Kohda *et al.* (1993)	A	48
						Koyama *et al.* (1993)●●	A	59
						Noble *et al* (1993)	A	48

N = number of sequences; *L* = alignment length; *C* = percentage conservation. SS shows where secondary structure definitions come from D = DSSP; A = author's assignments. ● denotes those observed accuracies taken from the literature: kinase accuracy reported by Thornton *et al.* (1991); SH3 domain accuracy reported by Rost and Sander (1992). ●● a 15 residue, 3 helix insertion was removed from this structure, since it is absent in the others, and not considered during a consensus prediction. The results of this study do not vary significantly if a different method of secondary structure assignment (Richards & Kundrot, 1988) is used (unpublished results).

A program to calculate *C* and the expected range of prediction accuracy is available from the authors (INTERNET: geoff@biop.ox.ac.uk).

The authors thank Professor L. N. Johnson for encouragement and support. R.B.R. is a Commonwealth Scholar and a member of Keble College, Oxford. G.J.B. thanks the Royal Society for support.

References

Barton, G. J. & Russell, R. B. (1993). Protein structure prediction. *Nature (London)*, **361**, 505–506.

Barton, G. J., Newman, R. H., Freemont, P. F. & Crumpton, M. J. (1991). Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur. J. Biochem.* **198**, 749–760.

Bazan, J. F. (1990). Structural design and molecular

maximum and minimum consensus. More information about the effect of different alignment lengths was obtained by splitting the 14 initial structural alignments into smaller alignments of length 50, 75, 100, 125, 150, 175 and 200. Only alignments of 5 sequences were considered. When more than 5 sequences were present within an alignment, all possible 5 membered sub-alignments were generated up to a maximum of 200 sub-alignments. For alignments with more than 200 sub-alignments, a random sample of 200 sub-alignments was considered. Secondary structure definitions were obtained using the method of Kabsch & Sander (1983; DSSP). The output from DSSP was converted into a 3-state (helix, beta, coil) summary (helix = DSSP H,G; beta = DSSP E; coil = DSSP not H,G,E). Three state agreement between a secondary structure assignment and a consensus (whether predicted or derived as in Fig. 1) can be obtained from the equation:

$$Q_3 = w_{\text{helix}} + w_{\text{beta}} + w_{\text{coil}}$$

where w_x is the 2 state (i.e. helix or not helix *etc.*) percentage accuracy of the consensus when compared to the assignment for x = helix, beta, or coil:

$$w_x = \frac{\text{No. residues predicted correctly as type x}}{\text{sequence length}} \times 100.$$

Q_{3ave} in the Figure is the average of Q_3 calculated for the maximum and minimum consensus.

evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6934–6938.

Benner, S. & Gerloff, D. (1991). Patterns of divergence in homologous proteins and tertiary structure. A prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme. Regul.* **31**, 121–181.

Benner, S. & Gerloff, D. (1993). Predicting the conformation of proteins: man versus machine. *FEBS Letters*, **325**, 29–33.

Benner, S., Cohen, M. A. & Gerloff, D. (1992). Correct structure prediction? *Nature (London)*, **359**, 781.

Benner, S., Badcoe, I., Cohen, M. & Gerloff, D. (1993). Predicted secondary structure for the *src* homology 3 domain. *J. Mol. Biol.* **229**, 295–305.

Berstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanovichi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bewley, M., Boustead, C., Walker, J., Waller, D. & Huber, R. (1993). Structure of chicken annexin V at 2·25 Å resolution. *Biochemistry*, **32**, 3923–3929.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature (London)*, **326**, 347–352.

Booker, G. W., Breeze, A. L., Downing, A. K., Panayotou, G., Gout, I., Waterfield, M. D. & Campbell, I. D. (1992). Structure of an SH2 domain of the p85α subunit of phosphatidylinositol-3-OH kinase. *Nature (London)*, **358**, 684–687.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol.* **47**, 45–148.

Crawford, I. P., Niermann, T. & Kirchner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins: Struct. Funct. Genet.* **2**, 118–129.

Eck, M., Shoelson, S. & Harrison, S. (1993). Recognition of a high-affinity phophotyrosyl peptide by the *src* homology-2 domain of p56 *lck*. *Nature (London)*, **362**, 87–91.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Holley, H. L. & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 152–156.

Huber, R., Berendes, R., Burger, A., Schneider, M., Karshikov, A., Hartmut, L., Romisch, J. & Paques, E. (1992). Crystal and molecular structure of human annexin V after refinement: implications for structure, membrane binding and ion channel formation of the annexin family of proteins. *J. Mol. Biol.* **223**, 683–704.

Hyde, C., Ahmed, S., Padlan, E., Miles, E. & Davies, D. (1988). Three-dimensional structure of the tryptophan synthase α_2/β_2 multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* **25**, 17857–17971.

Kabsch, W. & Sander, C. (1983). A dictionary of protein secondary structure. *Biopolymers*, **22**, 2577–2637.

Knighton, D., Zheng, J., Ten Eyck, L., Xuong, N., Taylor, S. & Sowadski, J. (1991). Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine mono-phosphate dependent protein kinase. *Science*, **253**, 407–414.

Kohda, D., Hatanaka, H., Odaka, M., Mandiyan, V., Ullrich, A., Schlessinger, J. & Inagaki, F. (1993). Solution structure of the SH3 domain of phospholipase c-gamma. *Cell*, **72**, 953–960.

Koyama, S., Yu, H., Dalgarno, D., Shin, T., Zydowsky, L. & Schreiber, S. (1993). Structure of the PI3K SH3 domain and analysis of the SH3 family. *Cell*, **72**, 945–952.

Lesk, A. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of globins. *J. Mol. Biol.* **136**, 225–270.

Lim, V. (1974). Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88**, 873–894.

Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *CABIOS*, In the press.

Musacchio, M., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). Crystal structure of a *src*-homology 3 (SH3) domain. *Nature (London)*, **359**, 851–855.

Noble, M., Musacchio, A., Saraste, M., Courtneidge, S. & Wierenga, R. (1993). Crystal structure of the SH3 domain in human fyn; comparison of the three-dimensional structure of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**, 2617–2624.

Overduin, M., Rios, C. B., Mayer, B. J., Baltimore, D. & Cowburn, D. (1992). Three dimensional solution structure of the *src* homology 2 domain of *c-abl*. *Cell*, **70**, 697–704.

Panayotou, G., Bax, B., Gout, I., Federwisch, M., Wroblowski, B., Dhand, R., Fry, M., Blundell, T., Wollmer, A. & Waterfield, M. (1992). Interaction of the p85 subunit of PI3-kinase and its N-terminal SH2 domain with PDGF receptor phosphorylation site: structural features and analysis of conformational changes. *EMBO J.* **11**, 4261–4272.

Richards, F. & Kundrot, C. (1988). Identification of structural motifs from protein co-ordinate data: secondary structure and first-level supersecondary structure. *Proteins: Struct. Funct. Genet.* **3**, 71–84.

Robson, B. & Garnier, J. (1993). Protein structure prediction. *Nature (London)*, **361**, 506.

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature (London)*, **360**, 540.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. Schneider, R. & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci.* **18**, 120–123.

Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.

Russell, R. B., J.Breed & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters*, **304**, 15–20.

Sternberg, M. J. E., Barton, G. J., Zvelebil, M. J. J. M., Cookson, A. J. & Coates, A. R. M. (1987). Prediction of antigenic determinants and secondary structures of the major aids virus proteins. *FEBS Letters*, **281**, 231–237.

Taylor, W. R. & Geisow, M. J. (1987). Predicted structure

for the calcium-dependent membrane-binding proteins p35, p36 and p32. *Protein Eng.* **1**, 183–187.
Taylor, W. & Thornton, J. (1983). Prediction of super-secondary structure in proteins. *Nature (London)*, **301**, 540–542.
Thornton, J., Flores, T., Jones, D. & Swindells, M. (1991). Prediction of progress at last. *Nature (London)*, **354**, 105–106.
Waksman, G., Kominos, D., Robertson, S., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., Resh, M. D., Rios, C. B., Silverman, L. & Kuriyan, J. (1992). Crystal structure of the phosphotyrosine recognition domain of SH2 of *v-src* complexed with tyrosine-phosphorylated peptides. *Nature (London)*, **358**, 646–653.
Weng, X., Luecke, H., Song, I., Kang, D., Kim, S.-H. & Huber, R. (1993). Crystal structure of human annexin 1 at 2·5 Å resolution. *Protein Sci.* **2**, 448–458.
Yu, H., Rosen, M., Shin, T., Seidel-Dugan, C., Brugge, J. & Schreiber, S. (1992). Solution structure of the SH3 binding domain of *src* and identification of its ligand-binding site. *Science*, **258**, 1665–1668.
Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by F. Cohen

Note added in proof. Since the acceptance of this manuscript, Drs B. Rost, C. Sander and Prof. S. A. Benner have kindly provided updated consensus predictions for the SH3 domains. The accuracies of these predictions (in the same order as in Table 1) are 75%, 73%, 61%, 60% and 80% for Rost and Sander, and 42%, 55%, 48%, 55% and 44% for Benner and Gerloff.