Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 7558–7562, August 1993
Biophysics

Improved prediction of protein secondary structure by use of sequence profiles and neural networks

(protein structure prediction/multiple sequence alignment)

BURKHARD ROST AND CHRIS SANDER

Protein Design Group, European Molecular Biology Laboratory, D-6900 Heidelberg, Germany

Communicated by Harold A. Scheraga, April 5, 1993

ABSTRACT **The explosive accumulation of protein sequences in the wake of large-scale sequencing projects is in stark contrast to the much slower experimental determination of protein structures. Improved methods of structure prediction from the gene sequence alone are therefore needed. Here, we report a substantial increase in both the accuracy and quality of secondary-structure predictions, using a neural-network algorithm. The main improvements come from the use of multiple sequence alignments (better overall accuracy), from "balanced training" (better prediction of β-strands), and from "structure context training" (better prediction of helix and strand lengths). This method, cross-validated on seven different test sets purged of sequence similarity to learning sets, achieves a three-state prediction accuracy of 69.7%, significantly better than previous methods. In addition, the predicted structures have a more realistic distribution of helix and strand segments. The predictions may be suitable for use in practice as a first estimate of the structural type of newly sequenced proteins.**

The problem of protein secondary-structure prediction by classical methods is usually set up in terms of the three structural states, α-helix, β-strand, and loop, assigned to each amino acid residue. Statistical and neural-network methods use a reduction of the data base of three-dimensional protein structures to a string of secondary-structure assignments. From this data base the rules of prediction are derived and then applied to a test set. For about the last 10 yr, three-state accuracy of good methods has hovered near 62–63%. Recently, values of 65–66% have been reported (1–4). However, when test sets contain proteins homologous to the learning set or when test results have not been multiply cross-validated, actual performance may be lower.

Point of Reference

We use as a "reference network" a straightforward neural-network architecture (5) trained and tested on a data base of 130 representative protein chains (6) of known structure, in which no two sequences have >25% identical residues. The three-state accuracy of this network, defined as the percentage of correctly predicted residues, is 61.7%. This value is lower than results obtained with similar networks (5, 7–10) for the following reasons. (*i*) Exclusion of homologous proteins is more stringent in our data base—i.e., test proteins may not have >30% identical residues to any protein in the training set. Other groups allow cross-homologies up to 49% [e.g., 2-hydroxyethylthiopapain (1ppd) and actinidin (2act) in the testing set termed "without homology" in ref. 5] or 46% (4). (*ii*) Accuracy was averaged over independent trials with seven distinct partitions of the 130 chains into learning and test set (7-fold cross-validation). The use of multiple cross-validation is an important technical detail in assessing performance, as accuracy can vary considerably, depending upon which set of proteins is chosen as the test set. For example, Salzberg and Cost (3) point out that the accuracy of 71.0% for the initial choice of test set drops to 65.1% "sustained" performance when multiple cross-validation is applied—i.e., when the results are averaged over several different test sets. We suggest the term sustained performance for results that have been multiply cross-validated. The importance of multiple cross-validation is underscored by the difference in accuracy of up to six percentage points between two test sets for the reference network (58.3–63.8%).

Use of Multiple Sequence Alignments

It is well known that homologous proteins have the same three-dimensional fold and approximately equal secondary structures down to a level of 25–30% identical residues (11). With appropriate cutoffs applied in a multiple sequence alignment (12), all structurally similar proteins can be grouped into a family, and the approximate structure of the family can be predicted, exploiting the fact that the multiple sequence alignment contains more information about the structure than a single sequence. The additional information comes from the fact that the pattern of residue substitutions reflects the family's protein fold. For example, substitution of a hydrophobic residue in the protein interior by a charged residue would tend to destabilize the structure. This effect has been exploited in model building by homology—e.g. in ref. 13—and in previous attempts to improve secondary-structure prediction (14–18). Our idea was to use multiple sequence alignments rather than single sequences as input to a neural network (Fig. 1). At the training stage, a data base of protein families aligned to proteins of known structure is used (Fig. 2). At the prediction stage, the data base of sequences is scanned for all homologues of the protein to be predicted, and the family profile of amino acid frequencies at each alignment position is fed into the network. The result is striking. On average, the sustained prediction accuracy increases by 6 percentage points. If single sequences rather than profiles are fed into a network trained on profiles, the advantage is generally lost.

Balanced Training

Most secondary-structure prediction methods have been optimized exclusively to yield a high overall accuracy. This method can lead to severe artifacts because of the very uneven distribution of secondary-structure types in globular proteins: 32% α-helix, 21% β-strand, and 47% loop (our data base). Usually, loops are predicted quite well, helices are predicted medium well, and strands are predicted rather poorly. This imbalance can be corrected if the network is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

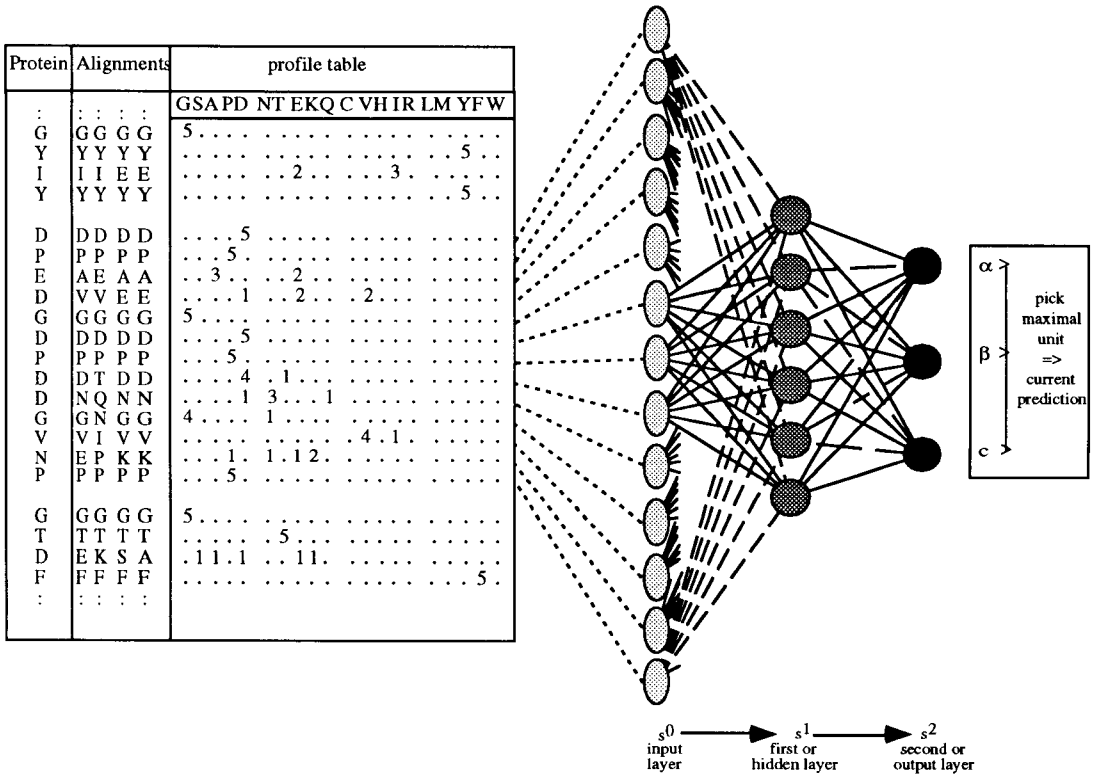

FIG. 1. Network architecture. A sequence profile of a protein family, rather than just a single sequence, is used as input to a neural network for structure prediction. Each sequence position is represented by the amino acid-residue frequencies derived from multiple sequence alignments as taken from the homology-derived structure of proteins (HSSP) data base (12). The residue frequencies for the 20-residue types are represented by 3 bits each (or by one real number). To code the N- and C-terminal ends an additional 3 bits are required (or one real number). The 63 bits originating from one sequence position are mapped onto 63 (21 for real numbers) input units of the neural network. A window of 13 sequence positions, thus, corresponds to 819 (273) input units. The input signal is propagated through a network with one input, one hidden, and one output layer. The output layer has three units corresponding to the three secondary-structure states, helix, β-strand, and "loop," at the central position of the input sequence window. Output values are between 0 and 1. The experimentally observed secondary structure states (19) are encoded as 1,0,0 for helix; 0,1,0 for strand; and 0,0,1 for loop. The error function to be minimized in training is the sum over the squared difference between current output and target output values. Net cascade: the first network (sequence-to-structure) is followed by a second network (structure-to-structure) to learn structural context (not shown). Input to the second network is the three output real numbers for helix, strand, and loop from the first network, plus a fourth spacer unit, for each position in a 17-residue window. From the $17 \times (3 + 1) = 68$ input nodes the signal is propagated via a hidden layer to three output nodes for helix, strand, and loop, as in the first network. In prediction mode, a 13-residue sequence window is presented to the network, and the secondary-structure state of the central residue is chosen, according to the output unit with the largest signal.

trained with each type of secondary structure in equal proportion (33%), rather than in the proportion present in the data base or anticipated in the proteins to be predicted. The result is a more balanced prediction (Fig. 3; Table 1), without affecting, negatively or positively, the overall three-state accuracy. A similar result was reported by Hayward and Collins (22). The main improvement is in a better β-strand prediction, the most difficult of the three states to predict. The method maintains full generality—i.e., it is equally applicable to all-α, mixed αβ, and all-β proteins. No knowledge of the structural type of the protein is required, as is the case for methods optimized on particular structural classes (9, 23).

Training on Structural Context

Even if a prediction method has high overall accuracy and is well balanced, it can be woefully inadequate in the length distribution of the predicted helices and strands. For example, the reference network predicts too many short strands and helices and too few long ones (Fig. 4). The predictions of this network appear fragmented compared with typical globular proteins. Published prediction methods have similar shortcomings in the length distribution of segments to various extents, except for two methods that optimize the sum of segment scores by dynamic programming (W. Kabsch and C.S., personal communication and ref. 24). The shortcoming is partly overcome here by feeding the three-state prediction output of the first, "sequence-to-structure," network, into a second, "structure-to-structure," network. The second network is trained to recognize the structural context of single-residue states, without reference to sequence information. Training it is very similar to that used for the sequence-to-structure network. The output string of the first network—e.g., the partially incorrect string HHHEEHH (two β-strand residues in the middle of a helix)—becomes the input to the second network and is confronted with correct structure HHHHHHH, a helical segment. Network couplings are optimized to minimize the discrepancy. The addition of the structure–structure network increases the overall accuracy only marginally but reproduces substantially better the length distribution of helices and strands. A simple way of measuring the quality of segment lengths is to compare the average length of helices and strands in the data base to those in the predicted structures ($\langle L_\alpha \rangle = 6.9$, $\langle L_\beta \rangle = 4.6$, Fig. 4). A similar second-level network was used by Qian and Sejnowski (5), but no effect of improved prediction of segment lengths was reported.

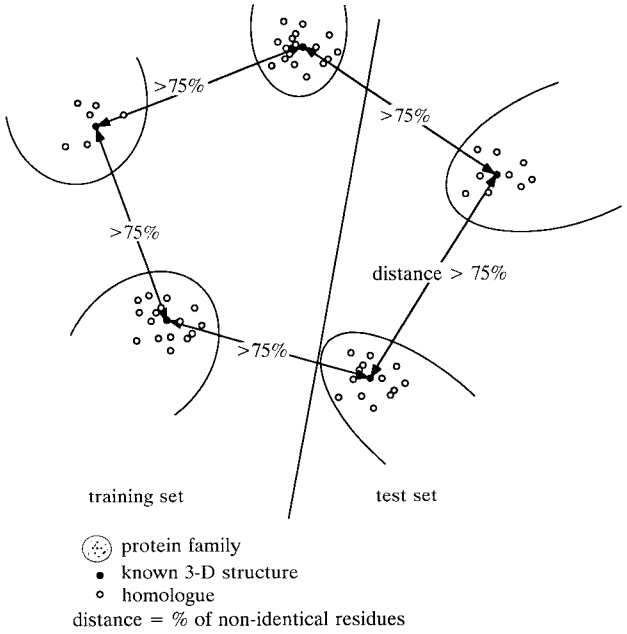

FIG. 2. Partition of protein families into training and test set. The structurally known representatives of the families used for training the network have a distance of at least 75% to those used for testing (sequence distance in percent nonidentical residues; drawn schematically). Each family contains homologous sequences, defined as those with a sequence identity >30% to the representative. 3D, three dimensional.

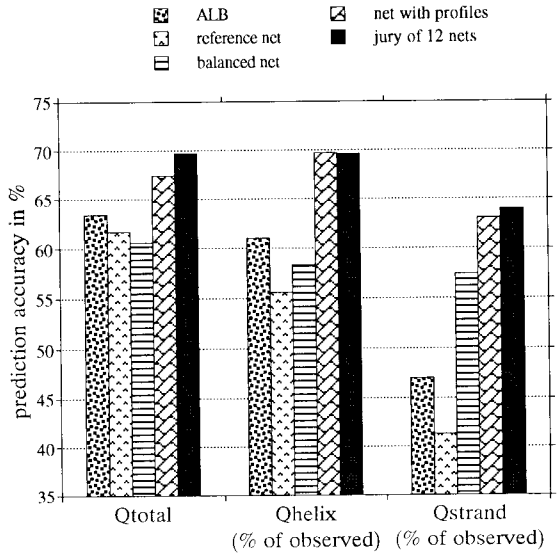

FIG. 3. Testing five secondary-structure prediction methods on the same set of proteins reveals the contribution of different devices to the improvement of accuracy. Q_{total}, overall prediction accuracy for the three states (helix, strand, loop; number of residues predicted correctly divided by the total number of residues). Q_{helix} and Q_{strand}, prediction accuracy calculated separately for helix and strand (e.g., number of helix residues predicted correctly divided by number of observed helix residues). The methods tested on our data base are ALB (20), first-level network with no balanced learning and no profiles (reference net), a two-level network cascade with balanced learning and no profiles (balanced net), a two-level network cascade with profiles and balanced learning (net with profiles), and 12 different networks combined by majority vote (jury of 12 nets). Some groups achieve higher accuracy than does ALB, but the accuracy values are not strictly comparable, as they are based on different test data sets and, in part, on test proteins with detectable sequence similarities to proteins on which the method was trained. Values for Q_{total} (Q_{helix}, Q_{strand}, Q_{loop}) are 65.5% (65, 45, 74), COMBINE (2); 63.0% (58, 54, 68), SIMPA (1); and 66.4%, Zhang *et al.* (4). Observed versus predicted matrix for the best method is indicated in Table 1.

"Jury of Networks"

An additional two percentage points in overall accuracy were gained by a jury of networks that predicts by simple majority vote of a set of 12 different networks. The increased accuracy is an effect of noise reduction, mitigating the ill effects of incomplete optimization when any single network settles into a local minimum of the error function.

Overall Improvement

The final jury of networks outperforms all known methods in overall accuracy, balanced β-strand prediction, and length distribution of segments as follows.

(*i*) The overall accuracy is 69.7%, three percentage points above the highest value reported so far [66.4% (4)]. The actual improvement may be larger, as their test set has sequence similarities of up to 46% relative to the training set. The improvement is six percentage points relative to the best classical method tested on our data base [63.4%, ALB (20)]. For a new protein sequence, one can expect a prediction accuracy between 61% and 79% (1 SD about the average over individual proteins of 70.2%), provided several homologous sequences are available. Values for three-state accuracy should not be confused with those for two-state accuracy (9, 23). Two-state predictions—e.g., for the state helix/nonhelix—carry less information and have a base value for random prediction of 50%—i.e., 17 percentage points higher than that for three-state methods.

(*ii*) Accuracy is well-balanced at 70% helix and 64% strand, measured as the percentage "correct of *observed*" (Fig. 3). The percentages "correct of predicted"—i.e., the probability of correct prediction, given a residue predicted in a particular state—are 72% helix and 57% strand.

(*iii*) The length distribution of segments is more "protein-like" (Fig. 4). Unfortunately, the length distribution is not

Table 1. Observed versus predicted matrix for best method of Fig. 3

	Residues predicted			Total observed	Residues predicted correctly,* %	
	Helix	Strand	Loop		Of observed	Of predicted
Residues observed						
Helix	5552	774	1,646	7,972	70	72
Strand	517	3229	1,310	5,056	64	57
Loop	1548	1592	8,227	11,367	72	73
Total predicted	7617	5595	11,183	24,395		
Correlation coefficient (21)	0.58	0.50	0.50			

*Prediction of jury of 12 nets method.

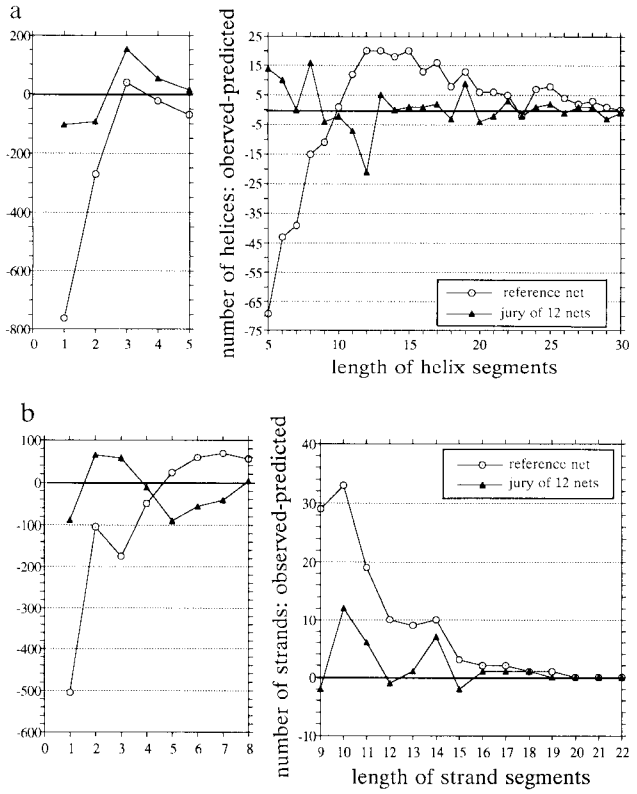

FIG. 4. Deviation in the length distribution of observed and predicted segments is an additional criterion by which prediction methods can be evaluated. (*a*) Difference in the length distribution of helix segments—i.e., number of observed segments in a given length range minus number of predicted segments. (*b*) Difference in the length distribution of strand segments. Predictions by the simple net (no profile, not balanced, no cascade) result in too many short segments, too few long segments; prediction by the jury of 12 nets results in a length distribution much closer to the observed one. Average segment lengths are as follows: reference net, $\langle L_\alpha \rangle = 4.2$ and $\langle L_\beta \rangle = 2.9$ residues; jury of 12, $\langle L_\alpha \rangle = 8.9$ and $\langle L_\beta \rangle = 5.1$ (observed: $\langle L_\alpha \rangle = 9.0$ and $\langle L_\beta \rangle = 5.1$).

generally given in the literature, but most methods are inferior in this regard.

Tests on Completely New Proteins

How accurate are predictions likely to be in practice? As a final check, the network system was trained on the full set of 151 sequence families of known structure and then tested on 26 protein families for which a first x-ray or NMR three-dimensional structure became available after the network architecture had been finalized. None of these additional test proteins had >25% sequence identity relative to any of the training proteins (Fig. 5). In this final set, 72% of the observed helical and 68% of strand residues were predicted correctly. The overall three-state accuracy for this set of completely new protein structures was 70.3%.

Predictions via Electronic Mail

Secondary-structure predictions using the currently best version of the profile network from Heidelberg (PHD) are available via electronic mail. Send a message containing the word "help" to PredictProtein@EMBL-Heidelberg.de. In practice, the predictions give a good first hypothesis of the structural properties of any newly sequenced water-soluble protein and may be an aid in the planning of point-mutation experiments and in the prediction of tertiary structure.

Conclusion

There are two important practical limitations: most of the advantage of the current method is lost when no sequence homologues are available; and the method in its current implementation is not valid for membrane proteins and other nonglobular or non-water-soluble proteins.

A major limitation in principle of the current method lies in its limited goal: secondary structure is a very reduced description of the complexities of three-dimensional structure and carries little information about protein function. However, as long as reliable prediction methods for protein three-dimensional structure and function are not available, secondary-structure predictions of improved quality are useful in practice—e.g., for the planning of point-mutation experiments, for the selection of antigenic peptides, or for identification of the structural class of a protein. Indeed, interest in the community is substantial: during 6 mo. since submission of this manuscript, >3,000 predictions for a wide variety of sequences have been requested and served via electronic mail.

Looking ahead, we would not be surprised to see increasingly successful use of evolutionary information in attempts to predict more complex aspects of protein structure and function. Sequence families grouped around one structure as well as structural superfamilies with common folds but divergent sequences (26, 27) contain a wealth of information not available 14 yr ago at the time of the first attempts at using homologous sequences for improved prediction (16). Having posed the puzzle of protein folding, evolution may hand us the key to its successful solution.

Note Added in Proof. Since the submission of this paper (April 1993) the method described has been improved further. By explicitly using

```
number      .........1.........2.........3.........4.........5.........6.........7.........8
sequence    AFDGTWKVDRNENYEKFMEKMGINVVKRKLGAHDNLKLTITQEGNKFTVKESSNFRNIDVVFELGVDFAYSLADGTELTG
observed       EEEEEEEEE HHHHHHH    HHHHHHH     EEEEEEE  EEEEEEE    EEEEEEE    EEEE      EEEE
predicted         EEEE   HHHHHHHHHHHHHHHHHHHH    EEEEEE  EEEEEEE  EEEEEEEEEE  EHHEE       EE

number      .........9.........0.........1.........2.........3.
sequence    TWTMEGNKLVGKFKRVDNGKELIAVREISGNELIQTYTYEGVEAKRIFKKE
observed    EEEEE  EEEEEEEE     EEEEEEEEE  EEEEEEEE  EEEEEEEEE
predicted   EEEE   HHEEEEEE     HHHHHHHHH    EEEEEEE EEEEEEE
```

FIG. 5. Example of prediction for a protein sequence by the currently best method. The β-barrel structure of intestinal fatty acid-binding protein has just become available through Protein Data Bank [code 1ifb (25)]. Prediction accuracy is 71.8%. In this β-sandwich structure, 8 out of the 10 β-strands are predicted correctly (one strand is ambiguous, and one strand is predicted as helix, but the ends of the segment are correct), and the two helices are predicted as one long helix (E: strand, H: helix). For all 26 new protein chains, including 1ifb, overall accuracy averaged over single residues is 70.3%; averaged over single proteins, it is 71.1%. The estimated probabilities of correct prediction, given a residue *predicted* in a helix, strand, or loop were 69%, 58%, or 77%, respectively (see text for probabilities relative to the number of residues *observed* in the three states). These 26 protein chains were not available publicly at the time of development of the method and were only used once in a final test of the currently best method. They are as follows: 1ace, 1cox, 1cpk_E, 1dfn_B, 5enl, 1f3g, 3fgf, 2gb1, 1gly, 1gmf_A, 1hcc, 1hdd_C, 2hip_B, 1ifb, 1msb_A, 1nsb_B, 5p21, 1pi2, 2pk4, 1rop_A, 1sar_A, 2scp_A, 1snv, 3trx, 3znf, 2zta_A (all taken from the Protein Data Bank prerelease of July 1992; membrane proteins and proteins with many metals or SS bridges were not considered).

conservation weights and the numbers of insertions and deletions in the multiple sequence alignments as input to the network system, the sustained overall three-state accuracy becomes 71.4% on the same data set used in this paper.

We thank Gerrit Vriend and Reinhard Schneider for stressing the importance of sequence profiles and segment lengths and Michael Scharf for general support; L. Philipson for reducing administrative load; and the Human Frontiers Science Program and the European Community Bridge Program for financial support.

1. Garnier, J. & Levin, J. M. (1991) *Comput. Appl. Biosci.* **7,** 133–142.
2. Levin, J. M. & Garnier, J. (1988) *Biochim. Biophys. Acta* **955,** 283–295.
3. Salzberg, S. & Cost, S. (1992) *J. Mol. Biol.* **227,** 371–374.
4. Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992) *J. Mol. Biol.* **225,** 1049–1063.
5. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202,** 865–884.
6. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) *Protein Sci.* **1,** 409–417.
7. Holley, H. L. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
8. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, L., Olsen, O. H. & Petersen, S. B. (1988) *FEBS Lett.* **241,** 223–228.
9. Kneller, D. G., Cohen, F. E. & Langridge, R. (1990) *J. Mol. Biol.* **214,** 171–182.
10. Stolorz, P., Lapedes, A. & Xia, Y. (1992) *J. Mol. Biol.* **225,** 363–377.
11. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5,** 823–826.
12. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
13. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990) *Proc. R. Soc. London B* **241,** 132–145.
14. Benner, S. A. & Gerloff, D. (1990) *Adv. Enzyme Regul.* **31,** 121–181.
15. Barton, G. J., Newman, R. H., Freemont, P. S. & Crumpton, M. J. (1991) *Eur. J. Biochem.* **198,** 749–760.
16. Maxfield, F. R. & Scheraga, H. A. (1979) *Biochemistry* **18,** 697–704.
17. Russell, R. B., Breed, J. & Barton, G. J. (1992) *FEBS Lett.* **304,** 15–20.
18. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **195,** 957–961.
19. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
20. Ptitsyn, O. B. & Finkelstein, A. V. (1983) *Biopolymers* **22,** 15–25.
21. Matthews, B. W. (1975) *Biochim. Biophys. Acta* **405,** 442–451.
22. Hayward, S. & Collins, J. F. (1992) *Proteins* **14,** 372–381.
23. Muggleton, S., King, R. D. & Sternberg, M. J. E. (1992) *Protein Eng.* **5,** 647–657.
24. Schneider, R. (1989) Diploma thesis (Dept. of Biology, Univ. Heidelberg, F.R.G.).
25. Sacchettini, J. C., Gordon, J. I. & Banaszak, L. J. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7736–7740.
26. Richardson, J. (1981) *Adv. Protein Chem.* **34,** 168–339.
27. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992) *Protein Sci.* **1,** 1691–1698.