Progress in protein structure prediction?

Burkhard Rost, Reinhard Schneider and Chris Sander

Prediction of protein secondary structure is an old problem and progress has been slow. Recently, spectacular success has been claimed in the blind prediction of the catalytic subunit of the cAMP-dependent protein kinase. When predictions in this and other test cases are assessed critically, some claims of prediction success turn out to be exaggerated, but a kernel of real progress remains: protein structure prediction can be improved substantially when a family of related sequences is available. Enough so that molecular biologists equipped with a new amino acid sequence and a multiple sequence alignment in hand may be tempted to test the new prediction methods.

KNOWN THREE-DIMENSIONAL STRUCTURES[1] (currently about 1000) are hopelessly outnumbered by known protein sequences (currently about 26 000)[2], and the gap is widening. The day when the sequences of 100 000 proteins will be known is not far off. So how can we close this sequence–structure gap? The problem is not a difficult one when we discover that a new sequence is homologous to a protein of known tertiary structure, as 'modeling by homology' is a well developed art. But for about five out of six new protein sequences, there is no detectable homologous structure in current databanks[3]. In these cases, we are in the same position that protein biochemists have faced for at least 20 years: predict protein structure from the amino acid sequence as best you can – until one day in the distant future your colleague down the hall solves the structure experimentally by X-ray crystallography or NMR spectroscopy.

Classical approaches with up to 63–65% accuracy

In its classical and simplest incarnation, the prediction problem is posed as that of predicting whether each residue in the protein forms part of an α-helix (α), a β-strand (β) or a loop, i.e. predicting secondary structure. A variety of methods have been brought to bear on this problem, using statistical information[4–11], physico-chemical properties[12,13], sequence patterns[14–20], multilayered artificial neural networks[21–25] and/or incorporating evolutionary information from sequence families[26–31]. We have compiled data on several of the prediction methods in use over the last ten years (see Box 1). The average per-residue accuracy on test cases of known structure has hovered near 62–63% with values of 65–66% reported in some cases.

A per-residue count is not the only means of assessing the accuracy of a particular prediction method. In certain cases, the number of α-helices and β-strands, or their placement, may be more important (see Box 2). However, any evaluation of the accuracy of a prediction method can be flawed if inadequate controls (cross-validations) are not included (see Box 1).

Blind tests are a simple trick that avoids any doubts about cross-validation controls. Crystallographers or NMR spectroscopists who are about to solve a structure circulate the protein sequence to predictors; the predictions are then compared to the experimental result. The first blind test of this type was organized by G. E. Schulz on adenylate kinase in 1977[32].

Deus ex machina?

Sudden excitement was generated recently when Benner and Gerloff[28] claimed very high accuracy for the prediction of a protein in a blind test[33]. They had taken an aligned set of sequences for catalytic subunits of protein kinases (Src homologues) and predicted the secondary structure. Soon thereafter, Knighton *et al.*[34] solved the crystal structure of the catalytic subunit of the cAMP-dependent protein kinase. The cystallographers' judgement was that the prediction was 'remarkably accurate, particularly for the small lobe' and the predictors' own assessment was 'the predictive power of the method developed in Zürich is far better than any that have preceded it', being based on 'a revolution in our understanding of the evolution of proteins and an important breakthrough in technology for organizing and analysing protein sequence data'.

These claims should be heard with caution. Success in this case does not necessarily guarantee future success. The Zürich method is carried out by hand using a combination of computer methods and human intuition, bers of proteins (as shown in Box 1) impossible. The actual per-residue prediction success of Benner *et al.* for the cAMP-dependent protein kinase is 63%, which is not particularly impressive. At least one fully automatic method does significantly better on this test case (76% for single residues in a blind test[35]).

Blind tests

However, blind tests are an excellent way of testing all prediction methods. Just in time for this review, the results of a second blind test became available: prediction and true structure of the Src-homology (SH3) domain of protein kinases were published back-to-back[36,37]. The result? Benner *et al.* achieve a per-residue accuracy of about 55% and predict the location and type of four of five β-strands (80%) correctly, using their computer-plus-intuition method. The fully automatic profile network method achieves a much higher 70% for single residues and also 80% for segment location.

These exercises in structure prediction set the stage. At the current rate of the order of 100 new experimental protein structure determinations per year, many more blind tests will become possible if crystallographers and NMR spectroscopists make known that they are about to solve a new structure. The new generation of methods can then be tested in the toughest possible way. At the end of 1993 we should know much better to what extent real progress has

B. Rost, R. Schneider and **C. Sander** are at EMBL, Meyerhofstrasse 1, 69 Heidelberg,

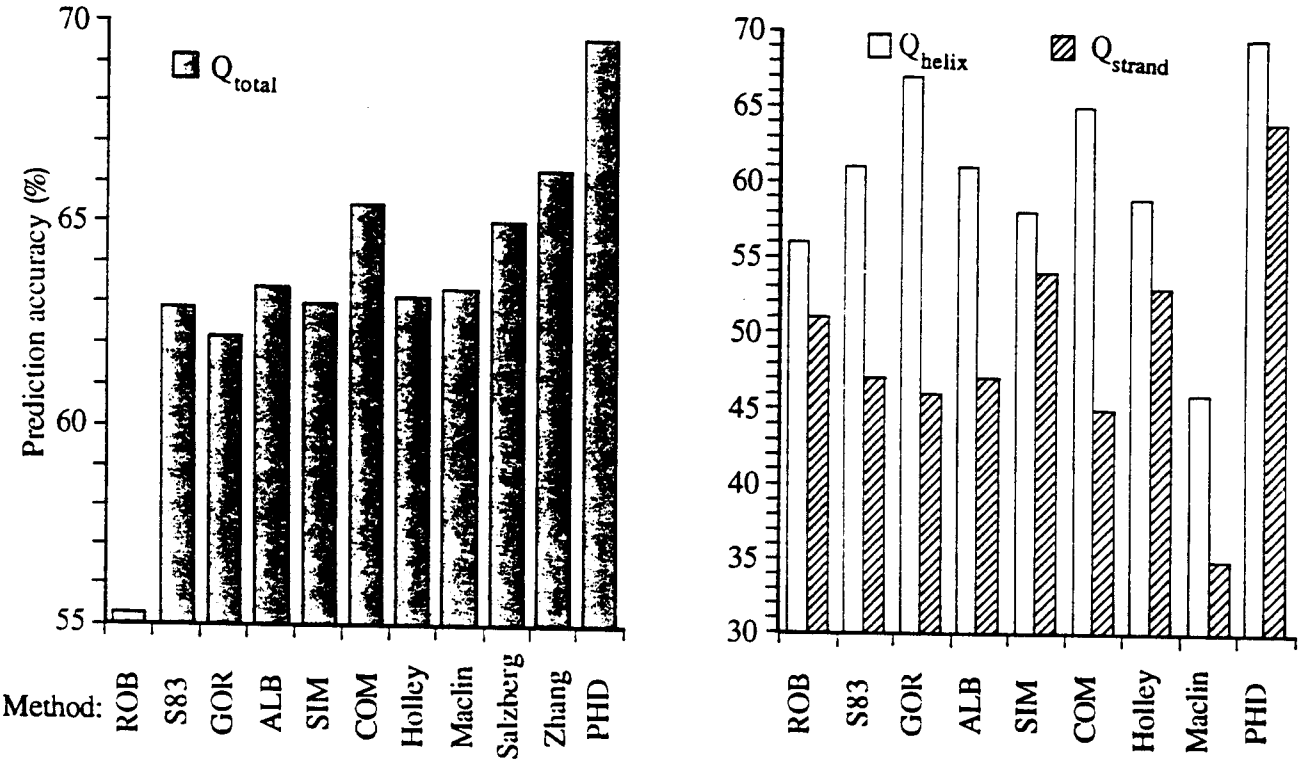
Box 1. How different secondary structure prediction methods compare when tested on many proteins

Measures of prediction accuracy: per-residue scores

A simple and classical way to measure accuracy of secondary structure prediction is to compare the experimental and predicted secondary structure, residue by residue. Normally this is done for the three states: α-helix, β-strand and loop. Q_{total} measures the overall accuracy and Q_{helix}, Q_{strand}, the accuracy for helices and strands.

$$Q_{total} = \frac{\text{number of correctly predicted residues}}{\text{total number of residues}}$$

$$Q_{helix\ (strand)} = \frac{\text{number of correctly predicted residues in helices (strands)}}{\text{number of residues observed in helices (strands)}}$$



The graphs above show the per-residue accuracy for nine selected methods. The accuracy measures are not strictly comparable, as different databases and different cross-validation methods are used. The methods ROB[6] (also known as GOR), S83 (C. Sander and W. Kabsch, unpublished), GOR[40] (known as GORIII), ALB[13], SIM[7] and PHD (Ref. 35 and B. Rost and C. Sander, submitted) were tested on a data set of about 150 protein chains with some 30 000 residues[41]. The values for COM[10] and Salzberg(92)[42] were taken from the literature, as were those for the multi-layered neural networks of Holley(89)[43], Maclin(92)[24], and Zhang(92)[25]. A randomly assigned prediction would, for example, yield a value of 36.3% for a typical data bank protein with 32% α, 21% β, 47% loop. The Q_{helix} and Q_{strand} accuracies were not available for all methods. Note: Holley did not cross-validate the predictions. Salzberg, Maclin, and Zhang have incomplete cross-validation, as there is significant sequence homology between some of the proteins used to derive the method and those used for testing its performance. Such homologies are a probable cause for overoptimistic expectations of performance. Some proteins are predicted less accurately than the database average and some more accurately. For example, the PHD method has an average Q_{total} of 69.7% (recently increased to 70.8%, B. Rost and C. Sander, unpublished) but accuracy for individual proteins varies between 45 and 95% with a standard deviation of 9.7%.

Proper evaluation?

Most prediction methods are derived from the database of known structures, but practice varies widely in how controls are performed. The most common errors and obfuscations in evaluating prediction methods are:

(1) evaluating the method on the same proteins that were used to derive the numerical parameters in the method;
(2) evaluating the method only on one particular division of the database into training and test set;
(3) using an unrepresentative or very small subset of proteins;
(4) evaluating the method on proteins with significant sequence similarity to the proteins used to derive the numerical parameters;
(5) choosing a measure of accuracy that looks particularly favorable.

We sought to avoid these errors in the profile network method (PHD) which uses a set of about 150 representative protein chains, purged of pairwise sequence similarities. Multiple cross-validation is performed by removing a subset (e.g. 15) of protein chains, training the network on the remaining 135 and finally applying the trained network to the 15 'unknown' chains. If this is done for ten different sets of 15 chains, then one can report the overall averages, Q_{total}, Q_{helix}, Q_{strand}, with accuracies that are 'tenfold cross-validated' for all 150 chains.

Box 2. Comparison of the experimental and predicted secondary structures of the catalytic subunit of the cAMP-dependent protein kinase

Measures of prediction accuracy: per-segment scores

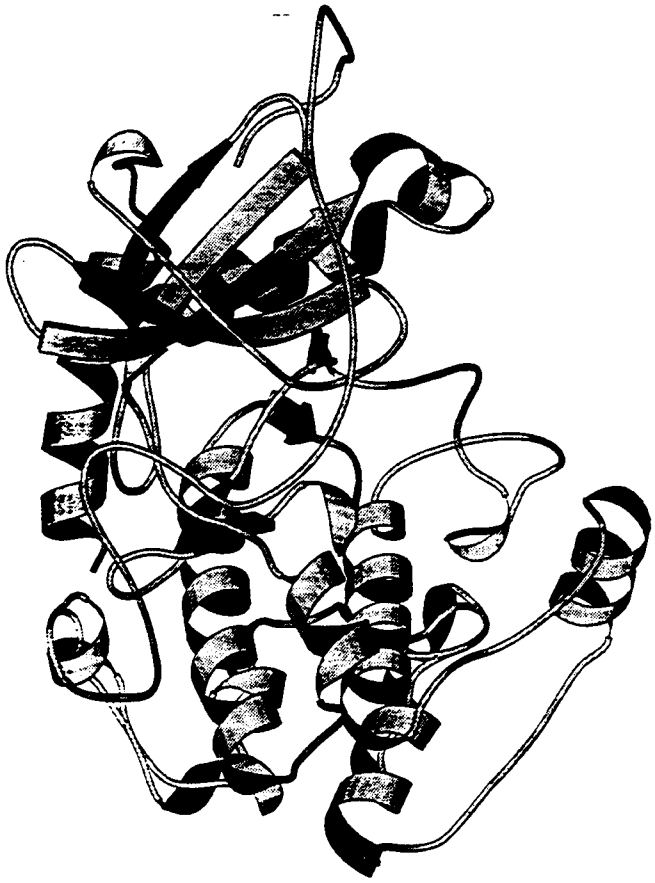
If the approximate placement of helices and strands is more important than a per-residue count, one can use a measure of per-segment accuracy. Making some allowance for errors at the end of segments, we propose that: an observed helix or strand is counted as correctly predicted if it overlaps with a predicted segment of the same type for at least half its length; an observed loop is counted as correctly predicted if at least two residues in it are predicted as loop.

Measures of prediction accuracy: segment lengths

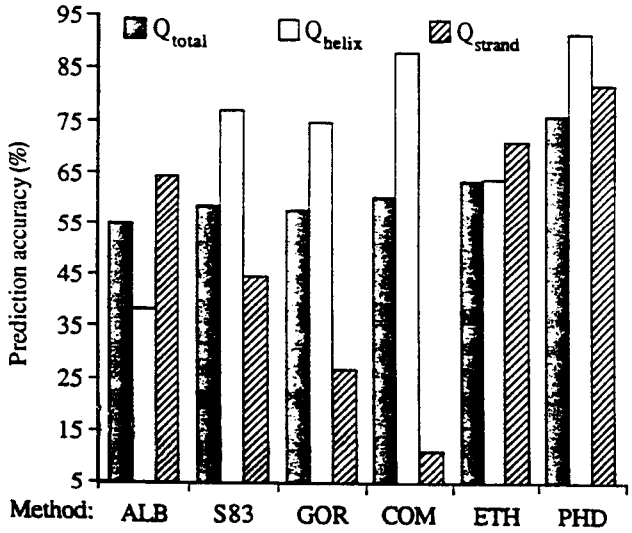
To predict the folding type of the protein (all-helical, (αβ)8-barrel, β-meander etc.) it is important not only to predict the correct placement of helices and strands, but also the correct number of segments.

```
     .........1.........2.........3.........4.........5.........6.
AA  |DQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNF|
OBS |HHEEEEEEEE        EEEEEE     EEEEEEEEHHHHHH    HHHHHHHHHHHH  |
COM |HHHHHHHH          EEEE        HHHHHHHHHHHHHHHHHHHHHHHHHHHHH  |
ETH |           EEEEEEEEEEEE      EEEEEEE            HHHHHHHHHHHHH  |
PHD |  HEEEEEEE     EEEEEEEE      EEEEEEEHHHHHHH    HHHHHHHHHHHH   |
     .........1.........2.........3.........4.........5.........6.
AA  |PFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL|
OBS |      EEEEEE    EEEEEEE       HHHHHHH     HHHHHHHHHHHHHHHHHH  |
COM |  HHHHHH          HHHHHHHH   HHHHHHHH        HHHHHHHHHHHHHHHHH |
ETH |  EEEEEEEEEE     EEEEEEEEE   HHHHHHHHH    HHHHHHHHHHHHHHHHHH |
PHD |   EEEEEEEEE     EEEEEEHH   HHHHHHHH      HHHHHHHHHHHHHHHH   |
     .........1.........2.........3.........4.........5.........6.
AA  |DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVD|
OBS |  EE         EEE      EEE       EE          HHH  HHHH        HH |
COM |HHHHHH   H  HHHH     EEE HHHHHH    EEEE          HHHH      HHH |
ETH |HEEEE       EEEE    EEEEEEEEE                EEEEEEEEEEE     EE |
PHD |  EEEE    HHHHHHHH    EEEEE  EEEE     EEEE      HHHHHHH     HHH |
     .........1.........2.........3.........4.........5.........6.
AA  |WWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKR|
OBS |HHHHHHHHHHHHH          HHHHHHHHHH          HHHHHHHHHH        |
COM |HHHHHHHHHHHH        H     HHHHHHH            HHHHHHHHHHHHHHHH |
ETH |EEEEEEEEEEEEE           HHHHHHHHHHH       HHHHHHHHHHHHHHH     |
PHD |HHHHHHHHHHHHH           HHHHHHHH          HHHHHHHHHHHHHH HHHH |
     .........1.........2.........3.........4.........5.........6.
AA  |FGNLKNGVNDIKNHKWFATT|
OBS |       HHHHH       |
COM |H       HHHHHHHHHHH|
ETH |  EEEEEEE          |
PHD |H       HHHHHHHHHHH|
```

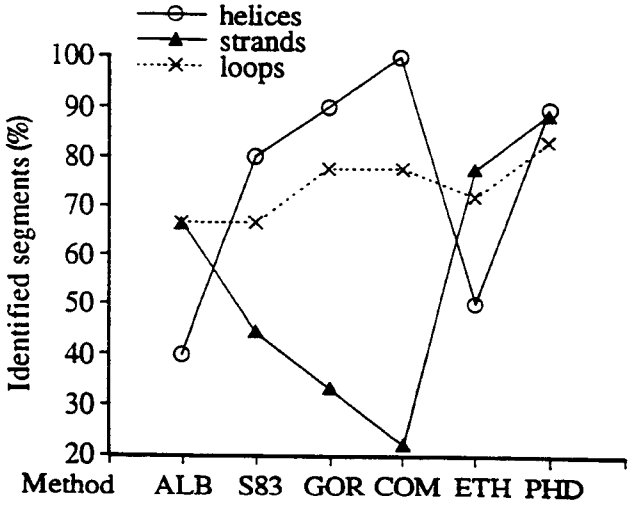
(a) Residue-by-residue comparison of experimentally observed (OBS) and predicted [COM[10], ETH[28], PHD (Ref. 35 and B. Rost and C. Sander, submitted)] structures of the catalytic subunit of the cAMP-dependent protein kinase (1cpk). 'AA' is the amino acid sequence taken from Protein Data Bank entry 1cpk (residues 27-287). Secondary structure: H = α-helix, E = β-sheet (extended), blank = loop. Predicted α-helices and β-strands that have insufficient overlap with an observed segment of the same type are underlined. Note the relatively good prediction of the location of segments for the ETH and PHD methods and overprediction of α-helices for the COM method.



(b) Ribbon view of the domain used in this blind test. The X-ray structure of catalytic subunit of the cAMP-dependent protein kinase. Drawn using Molscript[44].



(c) Per-residue prediction accuracy for 1cpk. The ETH and PHD methods give the overall best results.



(d) Per-segment prediction accuracy for 1cpk. Note the underprediction of α-helical segments by ETH and the 100% identification of α-helices at the expense of severe underprediction of β-strands in COM.

Evolutionary information is the key

Meanwhile, there are two important lessons. One is that several research groups have now demonstrated that the use of evolutionary information adds significantly to prediction success[29,30,33,35]. Maxfield and Scheraga were on the right track when they first used multiple sequence alignments to reduce statistical noise in 1979[26]. Barton *et al.* demonstrated this when they predicted the secondary structure of the SH2 phosphotyrosine binding domains and of annexin with remarkable accuracy[29,31].

What do multiple sequences tell you that single sequences don't? The sequence of a single protein encodes several functional messages, only one of which is spatial (secondary and tertiary) structure: other messages tell about the active site, protein–protein interactions, *in vivo* stability against degradation, membrane transport and so on. In addition, sequences contain mutational noise. So when information from an entire family of sequences, all of which have the same structure, is compounded and filtered, both noise and systematic error is reduced. What remains is a more clearly recognizable signal for structure formation. This effect is used profitably in *de novo* protein design[38,39]. When the effect is exploited for structure prediction of natural sequences, as in the profile network method[35], per-residue success is elevated to an average accuracy of 70%, cross-validated on more than 140 protein chains (see Box 1). But this is only the beginning. Predictors can exploit the rapidly growing database of protein sequences and sequence families and use evolutionary information in the form of multiple sequence alignments for a renewed attack on tertiary structure prediction. If the recent jump in the quality of secondary structure prediction is any indication, you will soon see more accurate predictions of aspects of tertiary structure.

To use or not to use?

In everyday practice, is it worth attempting to predict the secondary structure of each new gene sequence encountered? This depends on the question at hand. For example, those who plan mutational experiments to identify functional residues may find it a great time saver to know how to avoid mutations that simply destroy protein structure (e.g. a valine on the interior surface of a β-strand). Those looking for DNA-binding motifs may well benefit from knowing the location of safely predicted helices: when an appropriate cutoff is used, segments can be selected for which the expected accuracy is 90% or better. When caution is used and exaggerated expectations are circumvented, secondary structure prediction can indeed be a useful tool.

*The PHD method is available for fully automatic blind tests. Send the word 'help' by electronic mail to *PredictProtein@Embl-Heidelberg.de* for detailed instructions on how to automatically obtain the predicted secondary structure for your sequence.

References

1 Bernstein, F. C. *et al.* (1977) *J. Mol. Biol.* 112, 535–542
2 Bairoch, A. and Boeckmann, B. (1991) *Nucleic Acids Res.* 19, 2247–2250
3 Bork, P. *et al.* (1992) *Nature* 358, 287
4 Chou, P. Y. and Fasman, U. D. (1974) *Biochemistry* 13, 211–215
5 Nagano, K. and Hasegawa, K. (1975) *J. Mol. Biol.* 94, 257–281
6 Garnier, J., Osguthorpe, D. J. and Robson, B. (1978) *J. Mol. Biol.* 120, 97–120
7 Levin, J. M., Robson, B. and Garnier, J. (1986) *FEBS Lett.* 205, 303–308
8 Biou, V. *et al.* (1988) *Protein Eng.* 2, 185–191
9 Kanehisa, M. (1988) *Protein Eng.* 2, 87–92
10 Levin, J. M. and Garnier, J. (1988) *Biochim. Biophys. Acta* 955, 283–295
11 Fasman, G. F. (1989) in *The Development of the Prediction of Protein Structure* (Fasman, G. D., ed.), pp. 193–316, Plenum
12 Lim, V. I. (1974) *J. Mol. Biol.* 88, 857–872
13 Ptitsyn, O. B. and Finkelstein, A. V. (1983) *Biopolymers* 22, 15–25
14 Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. and Fletterick, R. J. (1983) *Biochemistry* 22, 4894–4904
15 Taylor, W. R. and Thornton, J. M. (1983) *Nature* 301, 540–542
16 Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. and Fletterick, R. J. (1986) *Biochemistry* 25, 266–275
17 Sternberg, M. J. E. and Islam, S. A. (1990) *Protein Eng.* 4, 125–131
18 Rooman, M. J., Kocher, J. P. and Wodak, S. J. (1991) *J. Mol. Biol.* 221, 961–979
19 Rooman, M. J. and Wodak, S. (1991) *Proteins* 9, 69–78
20 Presnell, S. R., Cohen, B. I. and Cohen, F. E. (1992) *Biochemistry* 31, 983–993
21 Bohr, H. *et al.* (1988) *FEBS Lett.* 241, 223–228
22 Qian, N. and Sejnowski, T. J. (1988) *J. Mol. Biol.* 202, 865–884
23 Kneller, D. G., Cohen, F. E. and Longridge, R. (1990) *J. Mol. Biol.* 214, 171–182
24 Maclin, R. and Shavlik, J. W. (1992) in *Refining Algorithms with Knowledge-Based Neural Networks: Improving the Chou–Fasman Algorithm for protein Folding* (Hanson, S. and Drostal, G. R., eds), MIT Press
25 Zhang, X., Mesirov, J. P. and Waltz, D. L. A. (1992) *J. Mol. Biol.* 225, 1049–1063
26 Maxfield, F. R. and Scheraga, H. A. (1979) *Biochemistry* 18, 697–704
27 Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987) *J. Mol. Biol.* 195, 957–961
28 Benner, S. A. and Gerloff, D. (1990) *Adv. Enz. Regul.* 31, 121–181
29 Barton, G. J., Newman, R. H., Freemont, P. S. and Crumpton, M. J. (1991) *Eur. J. Biochem.* 198, 749–760
30 Niermann, T. and Kirschner, K. (1991) *Protein Eng.* 4, 359–370
31 Russell, R. B., Breed, J. and Barton, G. J. (1992) *FEBS Lett.* 304, 15–20
32 Schulz, G. E. *et al.* (1974) *Nature* 250, 140–142
33 Benner, S. A. (1992) *Curr. Biol.* 2, 402–412
34 Knighton, D. R. *et al.* (1991) *Science* 253, 407–414
35 Rost, B. and Sander, C. (1993) *J. Neural Systems* (unpublished)
36 Benner, S. A., Cohen, M. A. and Gerloff, D. (1992) *Nature* 359, 781
37 Musacchio, A. *et al.* (1992) *Nature* 359, 851–855
38 DeGrada, W., Wassermann, Z. and Lear, J. (1989) *Science* 243, 622–628
39 Sander, C. (1990) *Biochem. Soc. Symp.* 57, 25–33
40 Gibrat, J-F., Garnier, J. and Robson, B. (1987) *J. Mol. Biol.* 198, 425–443
41 Altenberg, B. and Sander, C. (1992) *Current Quality of Secondary Structure Prediction*, EMBL
42 Salzberg, S. and Scott, C. (1992) *J. Mol. Biol.* 227, 371–374
43 Holley, H. L. and Karplus, M. (1989) *Proc. Natl Acad. Sci. USA* 86, 152–156
44 Kraulis, P. (1991) *J. Appl. Crystallogr.* 24, 946–950

TIBS reference lists

Authors of *TIBS* articles are asked to limit the number of references cited to provide non-specialist readers with a concise list for further reading. It is hoped that the citation of other, more extensive review articles rather than a comprehensive list of original articles enables interested readers to delve more immediately into the topic.