CABIOS

Vol.10. no.1. 1994
Pages 53–60

PHD—an automatic mail server for protein secondary structure prediction

Burkhard Rost, Chris Sander and Reinhard Schneider

Abstract

By the middle of 1993, >30 000 protein sequences had been listed. For 1000 of these, the three-dimensional (tertiary) structure has been experimentally solved. Another 7000 can be modelled by homology. For the remaining 21 000 sequences, secondary structure prediction provides a rough estimate of structural features. Predictions in three states range between 35% (random) and 88% (homology modelling) overall accuracy. Using information about evolutionary conservation as contained in multiple sequence alignments, the secondary structure of 4700 protein sequences was predicted by the automatic e-mail server PHD. For proteins with at least one known homologue, the method has an expected overall three-state accuracy of 71.4% for proteins with at least one known homologue (evaluated on 126 unique protein chains).

Introduction

The number of known protein sequences (30 000 SWISSPROT release 25.0; Bairoch and Boeckmann, 1992) is growing much faster than that of known protein structures (1000 PDB; Bernstein *et al.*, 1977). About 300 of the known structures are unique in terms of detectable sequence homology (Hobohm *et al.*, 1992; U.Hobohm, personal communication). This situation makes theoretical predictions of structural features of proteins increasingly necessary.

Suppose one has a sequence of unknown structure (SOS) and wants to know as much as possible about the structure. How can theory help? If there is a protein with a similar sequence to SOS in the data bank of known structures, model building by homology allows the prediction of the structure of SOS with reasonable accuracy (Greer, 1980, 1981, 1990, 1991; Blundell *et al.*, 1987; Taylor and Orengo, 1989; Overington *et al.*, 1990; Summers and Karplus, 1990; Vriend and Sander, 1991; Holm and Sander, 1992b; Levitt, 1992; Taylor, 1992). If not, i.e. if the SOS belongs to the majority of the 75% of known sequences which do not have homologues among the known three-dimensional structures (Schneider and Sander, 1993), there is still a chance to model the fold. If the SOS is very short, molecular dynamics could perhaps help to fold it up (Karplus and Petsko, 1990; Jernigan, 1992; Abagyan and Totrov, 1993; Dill, 1993). If the SOS is too long, there still is a chance of finding the three-dimensional structure: one can try to thread the SOS into a known structure, i.e. to find a protein of known structure which has no significant sequence similarity to SOS but is likely to have the same fold (Eisenberg and McLachlan, 1986; Baumann *et al.*, 1989; Overington *et al.*, 1990, 1992; Sippl, 1990; Crippen, 1991; Finkelstein and Reva, 1991; Lüthy *et al.*, 1991, 1992; Goldstein *et al.*, 1992; Holm and Sander, 1992a; Sippl and Weitckus, 1992; Ouzounis *et al.*, 1993; Stulz *et al.*, 1993). If this attempt also fails, prediction in three dimensions is no longer possible. One now has to revert to one dimension, i.e. to a prediction of one-dimensional strings of secondary structure assignment. More than 20 years of continuous effort to predict the secondary structure has led to the result that the performance is still far from 100% accurate; but 100% is not necessary. An important goal is that most secondary structure segments are predicted correctly. And for this there are promising methods.

EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

How accurate is secondary structure prediction?

A random prediction of secondary structure in three states (helix, strand, rest—here termed loop) yields an overall per-residue accuracy of 35% (Rost *et al.*, 1993). [Note: for two state predictions such as helix/non-helix the random value is ~55% (Rost and Sander, 1993b)]. This value provides a lower limit for the evaluation of predictions. Early methods such as those of Chou and Fasman (1974), Robson *et al.* (Garnier *et al.*, 1978; Robson and Pain, 1971) and Lim (1974) scored 14–19% above the random level (Kabsch and Sander, 1983b). In the 1980s the accuracy increased to ~60–66% (Ptitsyn and Finkelstein, 1983; Levin *et al.*, 1986; Gibrat *et al.*, 1987; Biou *et al.*, 1988; Gascuel and Golmard, 1988; Levin and Garnier, 1988; Salzberg and Cost, 1992; Zhang *et al.*, 1992), i.e. 24–30% above the random level (Figure 1). Using profiles from multiple sequence alignments a system of neural networks (dubbed PHD) was the first method to achieve a performance accuracy >70% when cross-validated on >100 unique proteins (Rost and Sander, 1993a), i.e. >34% above the random level (Figure 1). What is a reasonable goal for accuracy in secondary structure prediction? An upper limit for a very accurate prediction is given by the accuracy to be expected if a three-dimensional homologue to the SOS were known, i.e. if its three-dimensional structure can be modelled by homology with reasonable accuracy. A comparison of 140 protein pairs of known structure shows that these have ~88% of their residues in identical secondary structure states: helix, strand or loop (Rost *et al.*, 1994).

But how good is the prediction for the test case SOS? Of course, the answer can only be estimated from the statistics over tests with proteins of known three-dimensional structure. For the 126 proteins used in the cross-validation test of PHD the standard deviation of the latest version was 9.5%, i.e. the prediction of SOS is likely to be 71.4 ± 9.5% accurate. Secondary structure predictions are successful in capturing the clichés contained in the data bank. So, the more unusual the SOS protein is compared to known structures, the less likely is a good prediction. Two recent examples of prediction failure are the phosphatidylinositol 3-OH kinase p58_human (Kohda *et al.*, 1993; Koyama *et al.*, 1993) and the anti-freeze protein type III anpc_macam (Sönnichsen *et al.*, 1993): both recently solved structures and both predicted at low accuracy of ~40%. Thus, there is a small but non-vanishing chance that the prediction for SOS is grossly wrong. A more encouraging message is that the network prediction allows the identification of regions that are predicted with higher reliability. About 36% of all residues are predicted at an expected accuracy of 88%, i.e. comparable to what can be expected if homology modelling were possible for SOS (Figure 2).

How does PHD work?

The methods used to generate the automatic prediction of secondary structure by the PHD method are a profile alignment (algorithm: MaxHom/HSSP), which processes a profile of amino acid composition, and a system of neural networks that uses this profile for the prediction (PHD). The methods will be only briefly sketched here as they are described in more detail elsewhere (Schneider and Sander, 1991; Rost and Sander, 1993a).

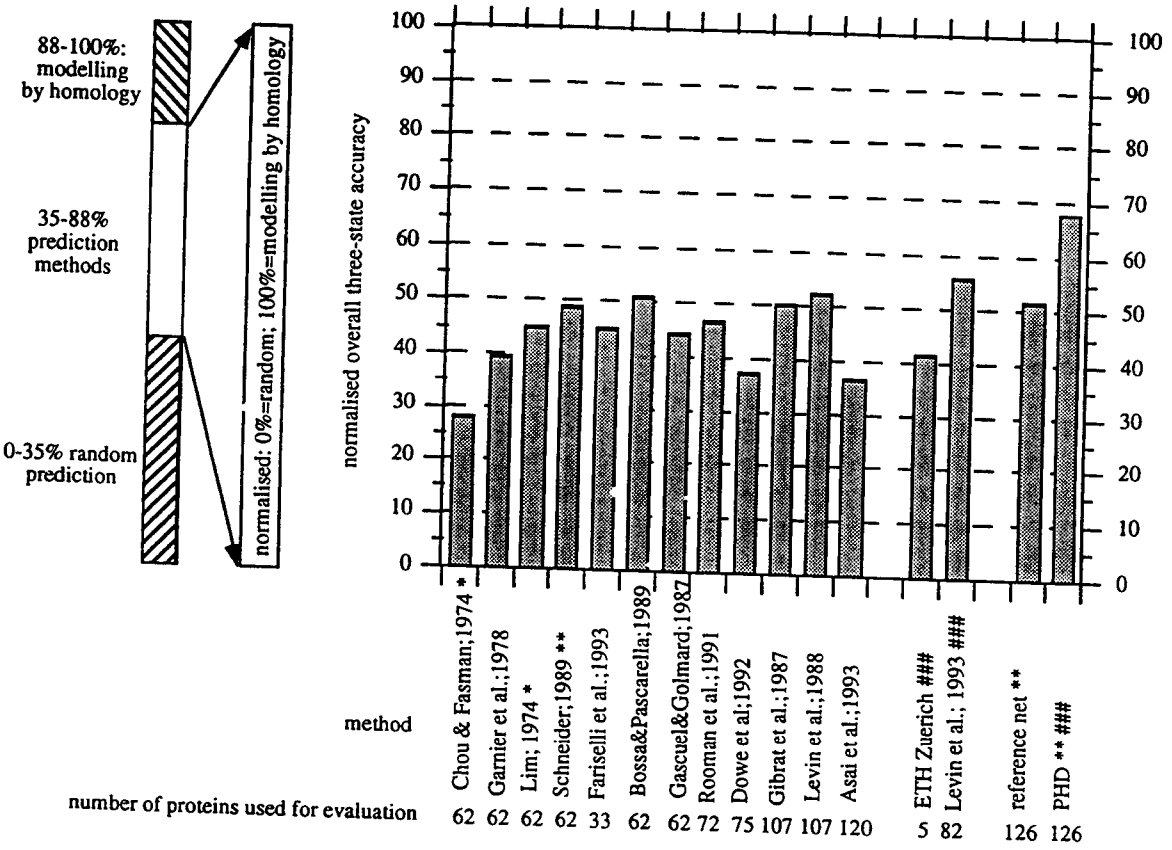

Fig. 1. Normalized overall three-state accuracy of secondary structure predictions. The values for the overall three-state (helix, strand, loop) accuracy of prediction methods are normalized such that a random prediction scores at 0%, and modelling by homology yields 100% [i.e. $Q_{plot} = Q_3^{method} - Q_3^{random})/(Q_3^{homology} - Q_3^{random})$]. Included are methods for which cross-validation has been performed without allowing pairwise sequence identity >25% between the proteins used for evaluation. The methods are labelled according to the citation in the literature list. Those labelled with a star (*) were tested on a database of 62 proteins used by Kabsch and Sander (1983b). A double star (**) indicates results for the database of 126 proteins used in Rost and Sander (1993a). 'Reference net' describes the performance of a standard neural network (Qian and Sejnowksi, 1988; Holley and Karplus, 1989) tested on 126 unique protein chains. 'PHD' labels the performance of the method used in the server. 'ETH Zurich' gives the result of the expert predictions by Benner *et al.* on five proteins (Benner and Gerloff, 1990, 1993; Benner *et al.*, 1993; Gerloff *et al.*, 1993). Levin *et al.* (1993) published a higher value of Q_3 = 69.6% for alignments of C^{α} traces. As this figure compares methods that predict secondary structure from the information available on the sequence level only, we used here the value of Q_3 = 64.9% the authors report for using multiple sequence alignments. For both the protein set predicted by 'ETH Zurich' and 'Levin et al., 1993' the PHD method scores above the average reported here (unpublished data). The methods marked with a triple hash ('# # # ') use multiple sequence alignments as input to the prediction.

HSSP—an alignment of multiple sequences by profiles

The MaxHom/HSSP algorithm builds up the alignment in essentially two steps. In sweep 1, the sequences are aligned consecutively to the guide sequence (SOS) by a standard dynamic programming method (Smith and Waterman, 1981).

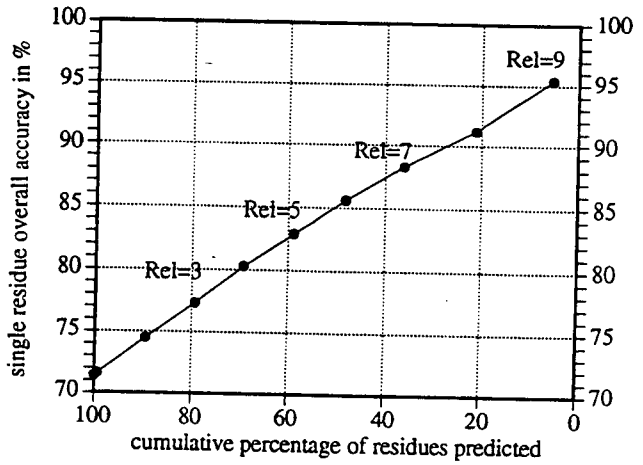

Fig. 2. Expected prediction accuracy for residues with a reliability index above a given cut-off. Plotted are averages of the three-state accuracy and the cumulative percentage of residues predicted over all those residues with reliability index (Rel) $\geq n$, $n = 0, \ldots, 9$. This index is simply defined by: Rel = INTEGER [10 * (out_{max} − out_{next})], where out_{max} is the output of the output unit with highest value, and out_{next} that of the unit with the next highest value. The factor 10 normalizes Rel to integer values from 0 to 9. Rel = 9 corresponds to a reliable prediction. For example, ~22% of all residues have Rel ≥ 8 and of these, 92% are correctly predicted by PHD.

After each sequence has been added to the alignment an alignment profile is compiled. This is used to align the next sequence. In sweep 2, after all sequences with significant homology have been picked from SWISSPROT, the profile is recompiled, and the dynamic programming algorithm starts once again to align consecutively the sequences, this time using the conservation profile as derived after completion of sweep 1.

PHD—a system of profile-reading neural networks to predict secondary structure

The profile and the conservation weight are used as input to a first level two-layered feed-forward network ('sequence-to-structure net'). This is done by shifting a window of 13 residues successively through the sequence, i.e. the nth window starts at position n and ends at $n + 13$ of the sequence. The output of the network consists of three values between 0 and 1, which give the probability that the residue in the centre of a particular window is in a helix, strand or loop. The first level sequence-to-structure net outputs the prediction for the central residue in a window. Thus, there is no direct correlation between the secondary structure of adjacent residues. This shortcoming is corrected by feeding the output of the first level sequence-to-structure net into a second level structure-to-structure network. The architecture of the second level network is the same as for the first level network. The third level is the computation of an arithmetic average over the outputs of several independently trained two-level nets (jury decision). The networks used for the third level (jury) are trained on the same training set, the

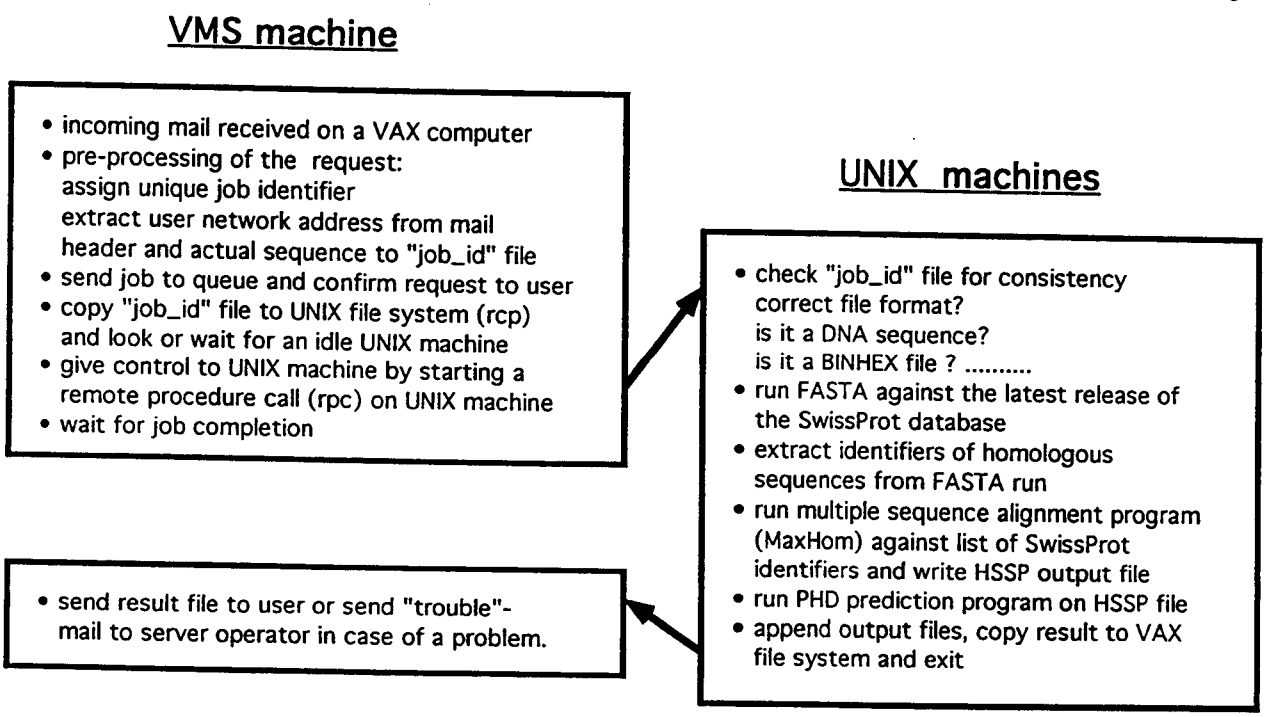

Fig. 3. Procedures performed by the PHD server. The Vax/VMS machine manages the incoming and outgoing mail, and sends the jobs to a cluster of four Unix machines. Here, the CPU-intensive processes are executed.

```
Joe Sequencer, Department of Advanced Protein Research,
National University, Timbuktu
joe@amino.churn.edu
# src homology-3 domain (SH3)

KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQ
GFVPAAYVKKLD
```

Fig. 4. Format of file to be sent to 'PredictProtein@EMBL-Heidelberg.DE'. Any format different from the one shown will result in an error message being sent back to the reader. The hash is necessary to recognize automatically that the sequence will start in the following line. If the server works fine, we do not look at the incoming prediction requests. Thus, messages to PredictProtein will remain unanswered. Instead, address queries or notes to Predict-Help@EMBL-Heidelberg.DE.

differences stem mainly from a different order of examples during the training procedure (Rost and Sander, 1993a).

PHD server—procedures between receiving and returning a mail

Communication between the VMS machine managing the incoming and outgoing mails, and Unix workstations used for the calculations is shown in Figure 3. The sequences are extracted from the mail, and the request is sent to a batch queue. From this queue, the jobs are sent to four Sun workstations (SPARC2 and 10). On the Unix side, the sequence is first checked for consistency. Then, the latest release of the SWISSPROT data bank of known sequences (Bairoch and Boeckmann, 1992) is searched for homologues. Currently, the system cuts off all sequences that have <30% sequence identity. This cut-off has been chosen as it is proven to yield true positives with high reliability. The majority of pairs in the 'twilight zone' between 25 and 30% sequence identity also have the same structure, but there also exist false positives in that region which would have to be filtered out by hand (Schneider and Sander, 1991). From the alignment the profile is computed and fed into the prediction program. Finally the result is copied (remote copy) to the VAX, and from there sent via electronic mail to the user.

How to obtain a PHD prediction

The sequence has to be written into a file according to the format shown in Figure 4. This file must be sent as electronic mail via internet to 'PredictProtein@EMBL-Heidelberg.DE'. For instructions one can send the word 'help' in the subject line to this address. Further questions, suggestions or notes should be addressed to 'Predict-help@EMBL-Heidelberg.DE'. A new option is that multiple sequence file formats (MSF; Devereux *et al.*, 1984) can be read, which means that the user can send his or her personal alignment for prediction. The alignment will then be used to compute a sequence profile in HSSP format to be used for the prediction.

```
## ALIGNMENTS    1 -   14
 SeqNo  PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  VAR
....:....1....:....2....:....3....:....4....:....5....:....6....
    1    6  K              0   0  184    5   24   KKK         E
    2    7  E            -  0   0   75   12   38   EEPETTT EDDR
    3    8  L  E         -AB 27  56A  71   12   43   LCQQILL YIIF
    4    9  V  E         -AB 26  55A   0   12   29   VVVAFFF VVVV
    5   10  L  E         -AB 25  54A  49   12   38   LVKRVVI RVVV
    6   11  A  E         - B  0  53A   2   13    0   AAAAAAAAAAAA
    7   12  L            +   0   0   55   13   12   LLLLLLLRLLLL
    8   13  Y  S       S-    0   0  126   13    1   YYYYYYYFYYF
    9   14  D            -   0   0   81   13   15   DDDDDDDDDPPD
   10   15  Y  B         -F 20  0B  18   13    2   YYYFYYYFYYY
   11   16  Q            -   0   0   94   13   41   QTDAEEECKDDA
   12   17  E            + -- 0   0   61   13   27   EEAAAAAAGGGA
   13   18  K            +   0   0  149   13   46   KKQERRRRNIIV
   14   19  S  S >  S-    0   0   32   13   43   SSTNTTTDDHHN
   15   20  P  T 3  S+    0   0  137   13   37   PPGPGEERDPPD
   16   21  R  T 3  S+    0   0  158   13   36   RRDDDDDSGDDR
   17   22  E  B <   -c  47  0A  10   13   11   EEEEDDDEDDDD
   18   23  V            -   0   0    2   13   11   VVLLLLLLLLLL
   19   24  T            +   0   0   55   13   33   TSTTTSTSPSSQ
   20   25  M  B         -F 10  0B   3   13   21   MMFFFFFLFFFV
   21   26  K     >      -   0   0  138   13   31   KKKNTRTKKKKL
   22   27  K  T 3  S+    0   0  137   13   19   KKEEKKKEKKKK
   23   28  G  T 3  S+    0   0   46   14    0   GGGGGGGGGGGG
   24   29  D     <      -   0   0   53   14   16   DDDAEEEDDEEED
   25   30  I  E         -A  5  0A  96   14   42   IVTVKRKIIKKKK
   26   31  L  E         -A  4  0A   0   14   28   LLIVFFFILMMLK
   27   32  T  E         -AD 3  40A  30   14   45   TTITHQHKKKKQD
   28   33  L  E         + D 0  39A  10   15   18   LLLVVIIIIIVVVL
   29   34  L  E         -   0   0   59   15   22   LLLHILLLLRLLLL
   30   35  N  E         + D 0  38A  72   15   34   NNNQNNNNNDEERF
   31   36  S            +   0   0   44   15   36   SSSKKNSNKKEESS
   32   37  T            +   0   0  127   14   40   TTNDSTTTKPHHT.
   33   38  N  S       S-    0   0   77   11   34   NNNPNEEEqE....
   34   39  K  S       S+    0   0  164   14   48   KKKAPYGGqEGGG.
   35   40  D  S       S+    0   0   77   14   18   DDDGDDDDDCQEED.
   36   41  W  E         - E 0  49A  76   15    1   WWWWWWWWWWWWF
   37   42  W  E         - E 0  48A  39   15    1   WWWWWWWWWWWWF
   38   43  K  E         +DE 30  47A  63   15   40   KKKEEEDERNKKLG
   39   44  V  E         -DE 28  46A   0   15   33   VVVGGAAAGAAAAV
   40   45  E  E         -DE 27  45A  57   15   32   EEEEERRREEkKre
   41   46  V  E >      - E 0  44A  16   15   41   VVVLLsssIDlsvv
   42   47  N  T 3  S-    0   0  149   15   44   NNNNNstsYMTsTD
   43   48  D  T 3  S+    0   0  112   15   27   DDDGGGGGGdKKGD
   44   49  R  E <   - E 0  41A 108   15   35   RRRKQHGKRkKRRL
   45   50  Q  E         + E 0  40A 102   15   43   QQQRRRSTVREEEQ
   46   51  G  E         - E 0  39A   2   15    8   GGGGGGGGGGGGV
   47   52  F  E         +cE 17  38A  56   15   24   FFFWVYYCWMFFYF
   48   53  V  E         - E 0  37A   0   15   17   VVVVFVIIFIIIVV
   49   54  P  E >      - E 0  36A  16   15    0   PPPPPPPPPPPPP
   50   55  A  G >  S+    0   0   15   15   32   AAAAASSSAVSSSP
   51   56  A  G 3  S+    0   0   76   15   34   AAANSNNNNPNNNA
   52   57  Y  G <  S+    0   0   89   15    1   YYYYYYYYYYYYF
   53   58  V  E <   -B  6  0A  11   15   10   VVIVVVVVVVVVVC
   54   59  K  E         -B  5  0A 107   15   39   KKKQEAAAEEAAAE
   55   60  K  E         -B  4  0A  82   15   43   KKKDLPPPEKKKPG
   56   61  L  E          B  3  0A  64   13   16   LLIIIVVV  LVVV
   57   62  D               0   0  193   12   29   DDD PDDD  NNEA
```

Fig. 5. Returned sequence alignment in HSSP format. Example for the SH3 protein (src homology region 3). The format of the multiple alignment is the same as the one used in the HSSP database of protein structure-sequence alignments. For the current PDB-SWISSPROT version the HSSP files are available via anonymous ftp from ftp.EMBL-Heidelberg.DE (Schneider and Sander, 1993). Abbreviations: SeqNo, numbering from first to last residue; PDBNo, position numbers from the related PDB file (column empty if there is no PDB file); AA, one-letter code for residue; STRUCTURE, secondary structure assignment according to DSSP (if three-dimensional structure is unknown this column gives a 'U' for unknown, *not* the predicted secondary structure); BP1 and BP2, positions of bridge partners in β-sheets; ACC, solvent accessibility as calculated by the DSSP program (Kabsch and Sander, 1983a); NOCC, number of sequences which are aligned at that position; VAR, residue type variability at that position. Lower-case characters in the alignment indicate deletions and dots mark insertions.

Soon after the sequence has been sent to the server, a message is sent to the user, confirming that the job has been sent to the queue. After the job has been processed, the user ought to receive the mail containing the output of the multiple alignment generated by the program MaxHom (Schneider and Sander, 1991) and written in the format of the HSSP files (Figure 5),

```
PHD output for your protein:
----------------------------

Abbreviations:
--------------

secondary structure : H=helix, E=extended (sheet), blank or L=rest (loop)
                 AA: amino acid sequence
                PHD: Profile network prediction HeiDelberg
                Rel: Reliability index of prediction (0-9)
detail:
                prH: 'probability' for assigning helix          --
                prE: 'probability' for assigning strand
                prL: 'probability' for assigning loop
               note: the 'probabilites' are scaled to the interval 0-9,
                     prH=5 means, that the signal at the first output no
                     is 0.5-0.6.
subset:
                SUB: a subset of the prediction, for all residues with
                     an expected accuracy > 82% (see  tables in header)
               note: for this subset the following symbols are used:
                  L: is loop (for which above " " is used)
                ".": means that no prediction is made for this residue,
                          as Rel < 5



          ....,....1....,....2....,....3....,....4....,....5....,....6
      AA  |KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD|
      Obs |  EEEE   E      E  E    EEEEEE     EEEEEE  EEEEEEHHEEEE  |
      PHD |  EEEEEEE         EE    EEEEEE      EEE      EEE   EEEE  |
      Rel |9468887622467997631215747999717873211243689842213111331 69
detail:
      prH-|0000000012210001110111010000010012333322110000111343111 00
      prE-|0378887742111000134532127888741011224552100035542124554 20
      prL-|8621011234567887754346761000147875431115678863345432234 78
subset: SUB |L.EEEEEE...LLLLLL....LL.EEEEE.LLL.......LLLL...........LL|
```

Fig. 6. Returned prediction of the PHD method. Example for the SH3 protein (src homology region 3). The PHD prediction is summarized in three lines: AA, echo of the sequence of SH3; PHD, prediction in three states helix (H), strand (E) and loop (blank); and Rel, the reliability index from 0 to 9, indicating a prediction site of highest reliability (definition as in Figure 2). Note: for comparison the DSSP (Kabsch and Sander, 1983a) assignment of the three-dimensional structure for SH3 (Musacchio *et al.*, 1992) is given (Obs). The second block of three lines dubbed 'detail' reports the probability of the assignment: the system of neural networks has three output units for helix, strand and loop, which can adopt values between 0 and 1. In the usual prediction (row 3: PHD) the highest of these units is chosen as the prediction. The detail gives the actual value for each unit (projected onto a grid from 0 to 9). This quantity supplies a probability for the assignment of helix, strand and loop. The last row (subset: SUB) repeats, for a quick overview, the same assignment as given in the third row (PHD), except that now only those assignments are given for which the prediction has an unexpected reliability index ≥ 5. For this subset the expected reliability is $>82\%$ (Figure 5).

and of the prediction of secondary structure in three states (Figure 6). Should >3 days elapse between the confirmation and the arrival of the prediction, we ask the user to write a note to Predict-Help@EMBL-Heidelberg.DE.

Recommendations after more than 4000 predictions

By September 15, 1993 >5500 predictions have been requested, and >4500 have been made (Figure 7). For ~ 1000 sequences no homologues were found in the sequence database. The number of requests varies considerably from one counry to another and is obviously affected by the size of the country, the level of activity in computational molecular biology, and the availability of molecular biology network services (Figure 8). For future requests, the user should keep the following points in mind.

Use homology prediction whenever possible

On occasion, the server alignment search turns up a homology to a protein of known three-dimensional structure. This is apparent from the presence of a four-letter PDB identifier (e.g. 5MBN for myoglobin) in the list of homologous proteins returned. In these cases, the predicted secondary structure gives, by definition, much less information than a three-dimensional model built using standard homology modelling tools.

Patterns not used for training cannot be predicted

The training of the network system was done on particular examples of native and dominantly globular proteins with 13 consecutive residues on the first level and 17 consecutive residues on the second level. Thus, the predictions cannot be expected to yield a comparable accuracy for:

- membrane proteins: average overall accuracy in three states was as low as 56% for porin (3por), melittin (2mlt), and the trans-membrane segments of the photo-reaction centre (1prc);
- fragments of <13 consecutive residues;
- differences between wild-type and point mutants;
- fragments containing a considerable percentage of unknown amino acids.

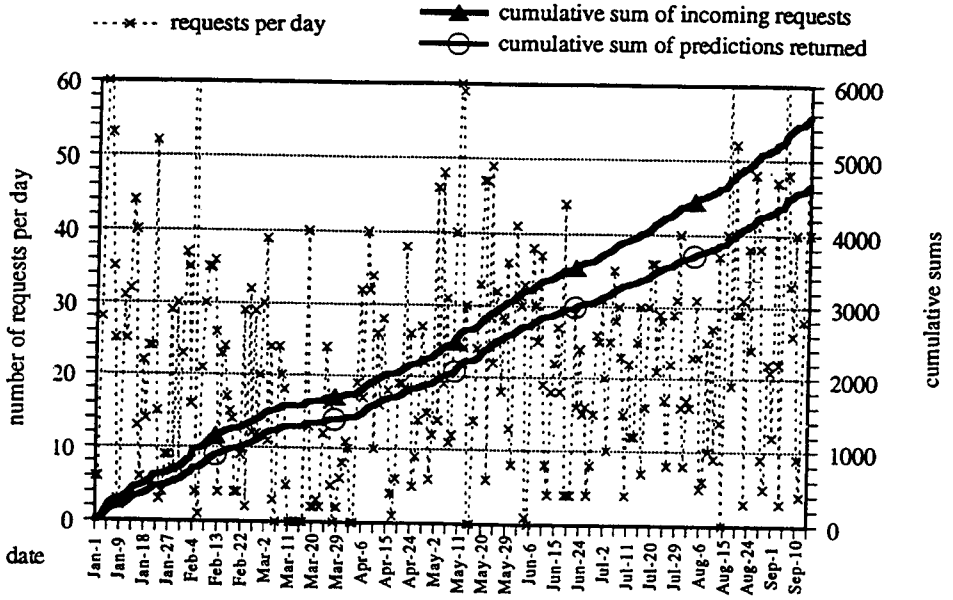

Fig. 7. Number of prediction requests from January to September 15, 1993. The difference between the requests and the number of predictions sent is partly explained by the fact that for the first four months no prediction was returned if no homologue to the sequence sent had been found in the current release of SWISSPROT. The decision to send predictions only if there was at least one homologue was motivated by the fact that only for this case has the network system been proven to be significantly better than alternative predictions. Meanwhile, we have lifted that restriction, as there is no evidence that the network without multiple alignment information is significantly worse than other prediction methods available [and is definitely much better than the Chou–Fasman method (Chou and Fasman, 1974)]. Expected accuracy figures, however, are for cases in which homologues are available. The server has become rather complex. This caused many difficulties, like the machine crash in the middle of March that led to a backlog of the requests.

The prediction depends on the quality of the multiple alignment

The accuracy of the prediction depends crucially on the information contained in the multiple alignment. PHD is significantly better than previously published methods only if a reasonable multiple sequence alignment can be made. Not only is the number of homologues important, but also the diversity of the family. It is better, in most cases, to have 50 sequences with 30–90% sequence identity relative to the guide sequence, than to have 100 sequences all in the range of 70–90% identity.

The prediction of the network, as given by the symbols H, E and L, can be rather sensitive to changes in the details of an alignment. However, this sensitivity does not hold for regions that are predicted with a high reliability index (Figures 2 and 6). Consequently, the prediction returned might be different for a different release of SWISSPROT for a given version of the trained networks. The probabilities for helix, strand and loop as given in the output of the returned prediction change only marginally for small deviations in the alignment profiles.

Cut-off in sequence similarity for the compilation of the multiple alignment

There is a length-dependent cut-off for structural similarity as a function of sequence similarity (Schneider and Sander, 1991). However, the cut-off is not razor sharp. Instead, there is a 'twilight zone' of, say, 3 percentage points (in sequence identity) above the cut-off line. The publicly available HSSP data bank

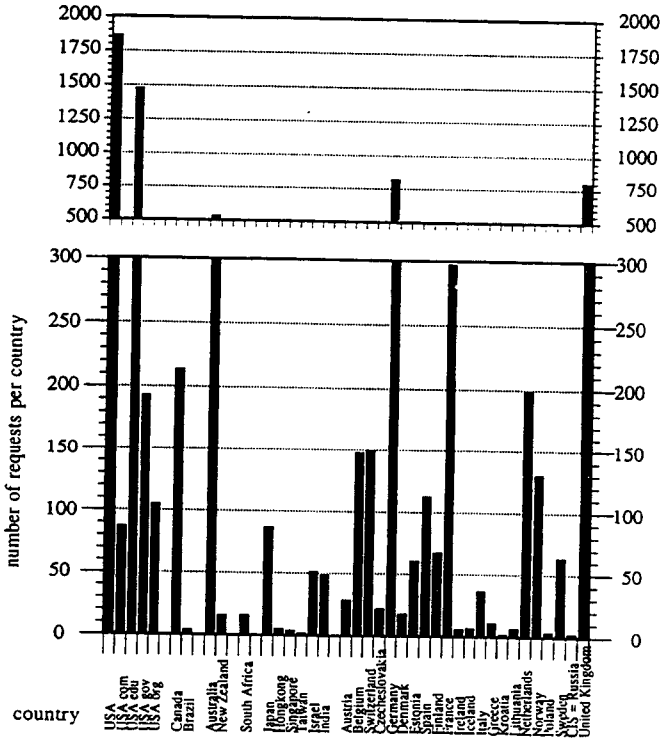

Fig. 8. Number of prediction requests per country. The level of server requests in a particular country reflects the number of molecular biology laboratories, the level of activity in computational molecular biology, the availability of network services or a combination of these and other factors.

uses the safety margin of 'cut-off + 5 percentage points', i.e. identical residues for alignments of length >80. Experiences with a set of 124 recently solved structures (B.Rost and C.Sander, unpublished results) have shown that the prediction accuracy is improved if the threshold is lowered to the threshold for structural homology defined earlier (Schneider and Sander, 1991). However, this may on occasion include spurious alignments, i.e. alignments with structurally non-similar proteins. Therefore, we currently use the safety margin of plus 5 percentage points for the server.

No consensus prediction

Although the network system uses the information contained in a sequence family, each prediction is derived for the secondary structure of the *guide* sequence. Strictly speaking, therefore, the result is not a consensus prediction for a family. A simple, practical way for every user to derive a consensus prediction is to predict the secondary structure for each sequence in a family, to bring the sequences in frame and then to sum up the probabilities for each output unit helix, strand and loop (given in the prediction output) at each alignment position over all predictions (in the sum, a weight reflecting the rarity of the sequence in the family should in principle be used). This is different from merely counting the predicted symbols at each position, an approach that would be inconsistent with our basic method. The last step is to assign the consensus prediction at each alignment position to the unit (helix, strand, loop) with the maximal sum. In practice, however, the difference between the prediction for the guide sequence and the consensus prediction for the entire family is marginal, so in most cases the server prediction can be safely taken as the family *consensus* prediction.

Uniqueness of the prediction

Since the service started, we have on three occasions changed the architecture of the networks used. The method is in the process of being improved further. Therefore, predictions sent at different times may differ not only as a result of database updates, but also as a result of updates in the method (see version number).

Conclusion

Over the last 20 years, some 100 methods for the prediction of protein secondary structure have been published. Only some of these methods are available to biologists who actually need the prediction tools. The PHD server gives access to a method that predicts secondary structure of globular proteins with an expected accuracy of 71.4%, if at least one homologue sequence can be aligned. The overall accuracy is some 5 percentage points better than any other method published (Zhang *et al.*, 1992). About 40% of all residues have an expected accuracy comparable to that achieved by homology modelling, i.e. modelling based on homology to a known three-dimensional structure. In comparison, for GORIII (Gibrat *et al.*, 1987) this value is reached for ~15% of all residues and for an alternative earlier neural network (Holley and Karplus, 1989) for ~10% of all residues. All the potential user needs is an electronic mail connection. This makes the service available also to users working in smaller laboratories without access to significant computer power. The speed of the prediction is sufficient to keep up with the amount of data produced by large-scale sequencing projects.

Acknowledgements

We are grateful to Roy Omond for implementing the server software managing the mail traffic and the internal communication between the VMS and the Unix cluster; to our colleagues Matthias Hage, Christos Ouzounis and Michael Scharf for having contributed software to the 'server package'; to the referee who helped with detailed and valuable remarks; and to all spectroscopists and crystallographers who made the experimental three-dimensional structures available. Last, not least, thanks to all users who helped with remarks, recommendations, often helpful questions, and their acknowledgement for what is sometimes a boring, laborious service. Finally, we want to apologize to all who suffered from hardware and software problems on our side.

References

Abagyan,R. and Totrov,M. (1993) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, **232**, in press.

Asai,K., Hayamizu,S. and Handa,K. (1993) Prediction of protein secondary structure by the hidden Markov model. *Comput. Applic. Biosci.*, **9**, 141–146.

Bairoch,A. and Boeckmann,B. (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **20**, 2019–2022.

Baumann,G., Frömmel,C. and Sander,C. (1989) Polarity as a criterion in protein design. *Prot. Engng*, **2**, 329–334.

Benner,S.A. and Gerloff,D. (1990) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.*, **31**, 121–181.

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

Benner,S.A., Cohen,M.A. and Gerloff,D. (1993) Predicted secondary structure for the Src homology 3 domain. *J. Mol. Biol.*, **229**, 295–305.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Biou,V., Gibrat,J.F., Levin,J.M., Robson,B. and Garnier,J. (1988) Secondary structure prediction: combination of three different methods. *Prot. Engng*, **2**, 185–191.

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

Bossa,F. and Pascarella,S. (1990) PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. *Comput. Applic. Biosci.*, **5**, 319–320.

Chou,P.Y. and Fasman,U.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 211–215.

Crippen,G.M. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232–4237.

Devereux,J., Haeberli,P. and Smithies,O. (1984) GCG package. *Nucleic Acids Res.*, **12**, 387–395.

Dill,K.A. (1993) Folding proteins: finding a needle in a haystack. *Curr. Opin. Struct. Biol.*, **3**, 99–103.

Dowe,D.L., Oliver,J., Dix,T.I., Allison,L. and Wallace,C.S. (1992) A decision graph explanation of protein secondary structure prediction. Department of Computer Science, Monash University, Australia.

Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Fariselli,P., Compiani,M. and Casadio,R. (1993) Predicting secondary structures

of membrane proteins with neural networks. *Eur. Biophys. J.*, **22**, 41–51.
Finkelstein,A.V. and Reva,B.A. (1991) A search for the most stable folds of protein chains. *Nature*, **351**, 497–499.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.
Gascuel,O. and Golmard,J.L. (1988) A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *Comput. Applic. Biosci.*, **4**, 357–365.
Gerloff,D.L., Jenny,T.F., Knecht,L.J., Gonnet,G.H. and Benner,S.A. (1993) The nitrogenase MoFe protein. *FEBS Lett.*, **318**, 118–124.
Gibrat,J.-F., Garnier,J. and Robson,B. (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, **198**, 425–443.
Goldstein,R.A., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992) Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. USA*, **89**, 9029–9033.
Greer,J. (1980) Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. USA*, **77**, 3393–3397.
Greer,J. (1981) Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.*, **153**, 1027–1042.
Greer,J. (1990) Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins*, **7**, 317–334.
Greer,J. (1991) Comparative modeling of homologous proteins. *Methods Enzymol.*, **202**, 239–252.
Hobohm,U, Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Prot. Sci.*, **1**, 409–417.
Holley,H.L. and Karplus,M. (1989) Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA*, **86**, 152–156.
Holm,L. and Sander,C. (1992a) Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.*, **225**, 93–105.
Holm,L. and Sander,C. (1992b) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins*, **14**, 213–223.
Jernigan,R.L. (1992) Protein folds. *Curr. Biol.*, **2**, 248–256.
Kabsch,W. and Sander,C. (1983a) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Kabsch,W. and Sander,C. (1983b) How good are predictions of protein secondary structure? *FEBS Lett.*, **155**, 179–182.
Karplus,M. and Petsko,G.A. (1990) Molecular dynamics simulations in biology. *Nature*, **347**, 631–639.
Kohda,D., Hatanaka,H., Odaka,M., Mandiyan,V., Ullrich,A., Schlessinger,J. and Inagaki,F. (1993) Solution structure of the SH3 domain of phospholipase C-γ. *Cell*, **72**, 953–960.
Koyama,S., Yu,H., Dalgarno,D.C., Shin,T.B., Zydowsky,L.D. and Schreiber,S.L. (1993) Structure of the PI3K SH3 domain and analysis of the SH3 family. *Cell*, **72**, 945–952.
Levin,J.M. and Garnier,J. (1988) Improvments in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys Acta*, **955**, 283–295.
Levin,J.M., Robson,B. and Garnier,J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**, 303–308.
Levin,J., Pascarella,S., Argos,P. and Garnier,J. (1993) Quantification of secondary structure prediction improvement using multiple alignments. *Prot. Engng*, **6**, 849–854.
Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.
Lim,V.I. (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, **88**, 857–872.
Lüthy,R., McLachlan,A.D. and Eisenberg,D. (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, **10**, 229–239.
Lüthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**.
Musacchio,A., Noble,M., Pauptit,R., Wierenga,R. and Saraste,M. (1992) Crystal structure of a Src-homology 3 (SH3) domain. *Nature*, **359**, 851–855.
Ouzounis,C., Sander,C., Scharf,M. and Schneider,R. (1993) Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.*, **232**, 805–825.
Overington,J., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond.*, **B241**, 132–145.
Overington,J., Donnelly,D., Johnson,M.S., Sali,A. and Blundell,T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Prot. Sci.*, **1**, 216–226.
Pitsyn,O.B. and Finkelstein,A.V. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15–25.
Qian,N. and Sejnowksi,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.
Robson,B. and Pain,R.H. (1971) Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.*, **58**, 237–259.
Rooman,M.J., Kocher,J.P. and Wodak,S.J. (1991) Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *J. Mol. Biol.*, **221**, 961–979.
Rost,B. and Sander,C. (1993a) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
Rost,B. and Sander,C. (1993b) Secondary structure prediction of all-helical proteins in two states. *Prot. Engng*, **6**, 831–836.
Rost,B., Schneider,R. and Sander,C. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, in press.
Salzberg,S. and Cost,S. (1992) Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.*, **227**, 371–374.
Schneider,R. (1989) Sekundärstrukturvorhersage von Proteinen unter Berücksichtigung von Tertiärstrukturaspekten. Diploma thesis, Department of Biology, University of Heidelberg.
Schneider,R. and Sander,C. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
Schneider,R. and Sander,C. (1993) The HSSP data base of protein structure-sequence alignment. *Nucleic Acids Res.*, **21**, 3105–3109.
Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. *J. Mol. Biol.*, **213**, 859–883.
Sippl,M.J. and Weitckus,S. (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, **13**, 258–271.
Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.
Sönnichsen,F.D., Sykes,B.D., Chao,H. and Davies,P.L. (1993) The nonhelical structure of antifreeze protein type III. *Science*, **259**, 1154–1157.
Stultz,C.M., White,J.V. and Smith,T.F. (1993) Structural analysis based on state-space modeling. *Prot. Sci.*, **2**, 305–314.
Summers.N.L. and Karplus,M. (1990) Modeling of globular proteins. *J. Mol. Biol.*, **216**, 991–1016.
Taylor,W. (1992) New paths from dead ends. *Nature*, **356**, 478–480.
Taylor,W.R. and Orengo,C.A. (1989) A holistic approach to protein structure alignment. *Prot. Engng*, **2**, 505–519.
Vriend,G. and Sander,C. (1991) Detection of common three-dimensional substructures in proteins. *Proteins*, **11**, 52–58.
Zhang,X., Mesirov,J.P. and Waltz,D.L. (1992) Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, **225**, 1049–1063.

Received on July 27, 1993; accepted on September 30, 1993