Bull. Inst. Chem. Res., Kyoto Univ., Vol. 66, No. 4, 1989

A Highly Reliable Prediction of Protein Secondary Structure

Yasuo KONISHI* and Ken NISHIKAWA**

Received April 18, 1988

Secondary structures (α-helix, β-strand or other structures) of proteins were predicted using five different prediction methods availalbe from the literature. A test application to 79 proteins of known structure showed that a highly reliable prediction (90 % accuracy) was obtained when all five methods predicted the same structure. However, the method predicted only 25 % of the residues in a particular protein. In order to increase the number of residues predicted without losing reliability, secondary structure predictions of homologous sequences were combined since the secondary structure of homologous proteins are usually almost identical. The National Biomedical Research Foundation Sequence Data Bank contains homologous sequences for 58 out of the set of 79 proteins of known structure (73 %). A test application to these 58 proteins led to prediction of the secondary structures of 48 % of the residues with 85 % accuracy. Thus, if homologous sequences are known, the secondary structure of about half (on average) of the residues in a protein is predictable with high accuracy.

KEY WORDS: Protein Secondary Structure Prediction/ Sequence Alignment/ Homologous Sequence

INTRODUCTION

Although many prediction methods have been reported, the accuracy of these methods lies around 60% for the three-state (α-helix, β-strand and other structures) model. These methods take into account the effect of neighboring residues, which includes short- and medium-range interactions, but not long-range interactions. Kabsch and Sander[1)] reported that an identical sequence of five residues can adopt different secondary structures when found in two different proteins. This implies that long-range interactions are one of the factors that determines the secondary structure of proteins in addition to the accepted contribution of short- and medium-range interactions. Thus, 60% reliability may be close to the limit of prediction schemes that use information from short- and medium-range interactions only.

Generally, these methods attempt to predict secondary structure for every residue in a protein; however, certain residues in a protein are predicted with more confidence, while others are not. Thus, if the prediction is limited to those residues for which there is confidence in the prediction scheme, the reliability of the prediction will be high. Nishikawa and Ooi[2)] have reported high reliability (70%) of prediction for those regions of proteins for which three different prediction methods concurred. This compares to around 60% reliability for individual prediction methods, although the regions predicted by the combined method account for less than half of the total residues. This indicates that prediction of secondary structure for regions in which

* 小西 康夫: Monsanto Co., 700 Chesterfield Village Parkway, St. Louis, Mo 63198, USA.
** 西川 建: Laboratory of Physical Chemistry of Enzyme, Institute for Chemical Research, Kyoto University, Uji 611.

many methods agree would be highly reliable, but such regions would be a small fraction of the total sequence.

Homologous proteins that are proteins with the same function, but from different species, e.g., horse, rat, pig, etc., have nearly identical secondary structures[3], although slight variations in their sequences give rise to slight variations in secondary structure predictions. Thus, we reasoned that combining the predicted structures for homologous proteins might increase the fraction of residues predicted by our combined method with a minimum loss of reliability. The advent of DNA sequencing techniques has increased the amount of data available on homologous sequences. The National Biomedical Research Foundation (NBRF) Sequence Data Bank contains homologous sequences for 58 (73%) out of the 79 proteins with known x-ray structures (Brookhaven Protein Data Bank).

In this report five different prediction methods[2,4-7] were compared and the secondary structures of the residues were said to be predicted only when all of the five methods agreed. The scheme was applied to each member of families of homologous proteins. The results for a homologous family of proteins were in turn combined to give highly reliable secondary structure predictuons for as many as half of the total residues.

METHODS

Computer Programs for the Secondary Structure Prediction

Five different secondary structure prediction mehtods[2,4-7] were used in this work. The computer programs of refs. 2, 4 and 5 are in our hands[8]. The computer programs of ref. 6 and 7 were obtained from Scheraga and Nagano, respectively.

Homologous Proteins

The program FASTP written by Lipman and Pearson[9] was used to identify homologous families of proteins from the NBRF Sequence Data Bank and to align the sequences within the homologous family. Only proteins with the same name, from different but somehow related sources were used as homologous proteins in combining the secondary structure predictions. For example, pronghorn pancreatic ribonuclease, giraffe pancreatic ribonuclease, red deer pancreatic ribonuclease, etc., were used as sequences homologous to bovine pancreatic ribonuclease in the prediction. Thus, homologous proteins with different names were not included.

Prediction of Secondary Structures

Prediction of secondary structure for a protein sequence was performed as follows: First, proteins homologous to the target protein were identified and aligned as described above. Next, the secondary structures for each these homologous proteins were predicted. The secondary structure (α-helix, β-strand or coil) of a residue in these proteins was considered to be predicted only when all of the five prediction methods[2,4-7] predicted the same secondary structure; otherwise, no predicted structure was assigned to a residue. Secondary structure predictions for the aligned homologous proteins were then combined; if a residue of any member of the family was predicted, then that prediction was assigned to that residue regardless of whether or not the residue could be predicted in the original target sequence. If different structures were predicted for the same residue in different members of the family, then no predicted structure was assigned to this residue.

Accuracy of Prediction

The 79 proteins (14,114 residues) selected by Nakashima[10] from the Protein Data Bank[11] were used to evaluate the secondary structure prediction methods. Although the crystal structures of these p.oteins are solved, the assignment of the secondary structures for each residue has some ambiguity, especially at the edges of the secondary structures. In this report, the assignments of the secondary structures by the contributing crystallographer listed in the Protein Data Bank and also by the method of Kabsch and Sander[12] were used. The prediction was evaluated as correct if the prediction fits the secondary structure from either of these assignments. The accuracy of the prediction was evaluated as the percent fraction of the residues predicted correctly in the total residues predicted (not the total residues of the protein).

RESULTS

During the evolution of a protein, the variation of the amino acid sequence has occurred more rapidly than the variation of the three dimensional structure in order to maintain the biological function of the protein. Consequently, homologous proteins have slightly different amino acid sequences, although their overall secondary structures are nearly identical[3]. This sequence variation with nearly identical secondary sturctures can result in variation in the prediction of secondary structure; namely, the secondary structure of a residue in a protein may not be predicted with confidence, while the secondary structure of the corresponding residue in the homologous protein may be predicted in more confidence due to the sequence variation around the residue. Table 1 shows an example of the prediction of bovine pancreatic ribonuclease (RNase) A. The amino acid sequences and the secondary structure predictions of bovine pancreatic RNase A and guinea pig pancreatic RNase A from the 25-th to 40-th residues are:

```
Bovine RNase A        Sequence     YCNQMMKSRNLTKDRC
                      Prediction   C

                                      *   * **
Guinea pig RNase A    Sequence     YCNEMMKKREMTKDRC
                      Prediction       HHHHHH
```

where asterisks denote differences between the sequences. Although the α-helix around residue 30 was not predicted for the sequence of bovine pancreatic RNase A, it was correctly predicted using the homologous sequence of guinea pig pancreatic RNase A. Without the availability of homologous proteins only 31 residues (25%) out of 124 residues were predicted with an accuracy of 90 % for bovine pancreatic RNase A. The addition of eight homologous RNase A sequences to the prediction increased the number of the residues predicted to 65 (52 %) out of 124 with the accuracy of 91 %.

Fig. 1 plots the accuracy of the prediction against the percent of residues predicted in a protein. The 21 proteins (total 5158 residues) out of 79 that had no homologous sequences in the NBRF had a low 25 % of residues predicted with an overall accuracy of 90 % (filled triangles in Fig. 1). Similarly, when the secondary structures of the 79 proteins were predicted without grouping by homology, 25 % of the residues were predicted with an overall accuracy of 90 % (89 % for α-helix, 83 % for β-strand and 91 % for coil). When predictions of homologous sequences were combined for 58 proteins (total 8954 residues), the number of residues predicted was

Table I. Secondary Structure Prediction of Bovine Pancreatic Ribonuclease A.

The secondary structures (α-helix; H, β-strand; B and coil; C) are shown under the amino acid sequence; a blank space indicates that no structure was predicted for the residue. The protein sequences were obtained from the NBRF sequence data bank.
The number of total residues: 124 residues
The number of residues predicted: 65 residues (52%)
The number of residues predicted correctly: 59 residues
The number of residues predicted incorrectly: 6 residues
The accuracy of the prediction: 91%

```
Res. #          1 - 10    11 - 20    21 - 30    31 - 40    41 - 50
SEQUENCE     KETAAAKFER QHMDSSTSAA SSSNYCNQMM KSRNLTKDRC KPVNTFVHES
       a)
X-ray 1      CCHHHHHHHH HHHCCCCCCC CCCHHHHHHH HHHHCCCCCC BBBBBBBBCH
       b)
X-ray 2      CCCHHHHHHH HHCCCCCCCC CCCCHHHHHH HHCCCCCCCC CCCBBBCCCC
          c)
Prediction   CHHHHHHHH      CCCCCC CCCCCCC HH HHH CCCCC

    d)
nrbo          HHHHHHH       CCCCCC CCCCC
nrprh         HHHHHHHH      CCCC    CCCC            C
nrgf          HHHHHHH       CC      CCC           C CCC
nrder         HHHHHHHH      CCCCCC CCCCCCC           C
nrhp                        CCCCCC CCCCCC           CCC
nrgpa                       CCCCCC CCCC    HH HHHH   CC
nrbos         HHHHHHHH      CCCCCC CCCCC
nrwhk        C              CCCCCC CCCCCC
nrpg                        CCCCCC CCCC          CCCC


Res. #         51 - 60    61 - 70    71 - 80    81 - 90    91 - 100
SEQUENCE     LADVQAVCSQ KNVACKNGQT NCYQSYSTMS ITDCRETGSS KYPNCAYKTT
X-ray 1      HHHHHHHHHH BBBBCCCCCC BBBBBCCCBB BBBBBBBCCB BCCBBBBBBB
X-ray 2      HHHHHHHHHC BBBCCCCCCC CBBBCCCBBB BBBBCCCCCC CCCCCCBBBB
Prediction                   CCC      CCCCC B B   CCCCCC CCCCCC

nrbo                          CC
nrprh                        .CC
nrgf                         CCC        CC        CCCCCC CCCC
nrder                        CCC        C           CCCC CC
nrhp                         CC         C            CCC CCCC
nrgpa                        CC         CCC                 CCC
nrbos                                             CCCCCC CCCC
nrwhk                         CC      CCC     B             CC
nrpg                         CCC        C   B B
```

Table I. contioued

Res. #	101 - 110	110 - 120	121 - 124
SEQUENCE	QANKHIIVAC	EGNPYVPVHF	DASV
X-ray 1	BBBBBBBBBB	BBBBBBBBBC	BBBB
X-ray 2	BBBCCBBBBB	CCCCBBBBBB	BCCC
Prediction	BBBBBB	C	CCCC
nrbo	BBBBBB	C	CCCC
nrprh	BBB	C	
nrgf	BBBB	C	
nrder	BBBB	C	CCCC
nrhp	BBB		
nrgpa	BBBB		CCCC
nrbos	BBBB		CCCC
nrwhk	BBBB	C	CCCC
nrpg	BBBB		CCC

a) Secondary structures were assigned by crystallographers. Structures other than α-helix and β-strand were assigned as coil.

b) Secondary structures were assigned by the method of Kabsch and Sander[12]. Structures other than α-helix and β-strand were assigned as coil.

c) Secondary structure prediction on the basis of combined predictions for homologous protein sequences.

d) Secondary structure prediction of the individual sequences homologous to bovine pancreatic ribonuclease A using the five combined prediction schemes. The proteins are:

nrbo; Ribonuclease (EC 3.1.27.5), pancreatic - Bovine and American bison
nrprh; Ribonuclease (EC 3.1.27.5), pancreatic - Pronghorn
nrgf; Ribonuclease (EC 3.1.27.5), pancreatic - Giraffe
nrder; Ribonuclease (EC 3.1.27.5), pancreatic - Red deer and roe deer
nrhp; Ribonuclease (EC 3.1.27.5), pancreatic - Hippopotamus
nrgpa; Ribonuclease (EC 3.1.27.5) A, pancreatic - Guinea pig
nrbos; Ribonuclease, seminal (EC 3.1.27.-), α and β chains - Bovine
nrwhk; Ribonclease (EC 3.1.27.5), pancreatic - Minke whale
nrpg; Ribonclease (EC 3.1.27.5), pancreatic - Pig.

substantially increased to 48 % with a minimal loss of overall accuracy to 85 % (filled circles in Fig. 1). Since no correlation between the fraction of the residues predicted and the accuracy of the prediction is observed (Fig. 1), it is concluded that the prediction of homologous proteins can be combined to increase the number of the residues predicted with minimal loss of reliability in the prediction.

DISCUSSION

Although accurate secondary structural prediction is a difficult task, an old Chinese proverb states that "two heads are better than one". When secondary structures are predicted by only one of the available methods, the accuracy of the predictions is around 60 %. When five different prediction methods were combined, an interesting feature came out; the reliability of the prediction was very high (89 % for α-helix, 83 % for β-strand and 91 % for coil for 79 proteins without combining the prediction of homologous proteins) only when all of the five viewpoints predict the same structure. If one or two of them predicted a structure different from the others, i.e., a majority of the methods predict the same structure, the reliability of the

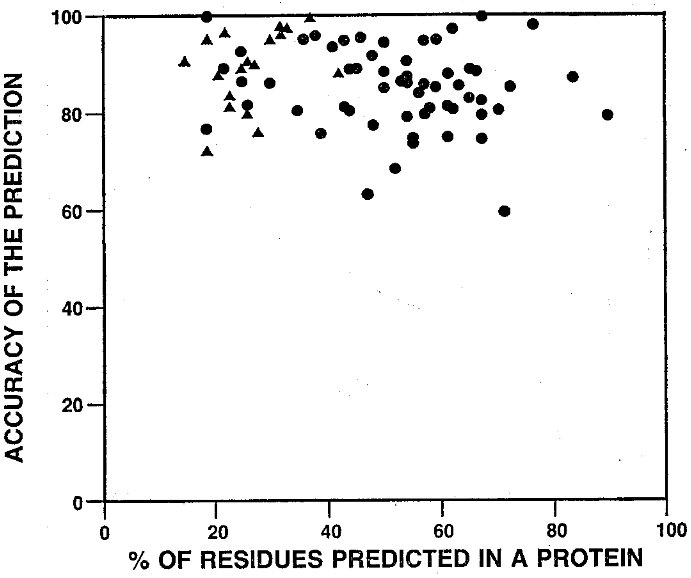

Fig. 1. The accuracy of the secondary structure prediction. 79 proteins of known structure were predicted. The 21 proteins (total 5158 residues) out of 79 had no homologous sequences in the NBRF. A low 25% of residues in these proteins were predicted the secondary structures with an overall accuracy of 90% (filled triangles) because of the lack of the homologous sequence information. The 58 proteins (total 8954 residues) out of 79 had homologous sequences in the NBRF. A high 48% of residues in these proteins were predicted the secondary structures with an overall accuracy of 85% (filled circles) when the prediction of homologous sequences were combined.

prediction is substantially lower (65 % for α-helix, 58 % for β-strand and 47 % for coil for 79 proteins without combining the prediction for homologous proteins). This means that democracy does not work in the predictions and unanimity is the way to predict secondary structures of proteins with high reliability.

Homologous proteins used were limited to ones with the same name from different but somehow related sources; in this report the cut-off of the homology was arbitrary, depending on the availability of the homologous sequence data. Less homologous proteins can be used to increase the number of residues predicted. These are the proteins with the same name, but not close in evolution or are different proteins for which homologous structures are expected, e.g., α-lactalbumin and hen egg lysozyme[13]. However, the accuracy of the prediction will be lower when less homologous proteins are used in the prediction. Fig. 2 shows the effect of homologous proteins on the accuracy of prediction and the number of residues predicted. The X-axis is the number of proteins homologous to bovine pancreatic RNase A used in the prediction. The most homologous proteins were included in the prediction first and then less homologous proteins were added later. Thus, proteins less homologous to bovine pancreatic RNase A are included in the prediction as the value on the X-axis increases. When only eight highly homologous proteins in which over 74 % of the residues were identical to bovine pancreatic ribonclease A, were included

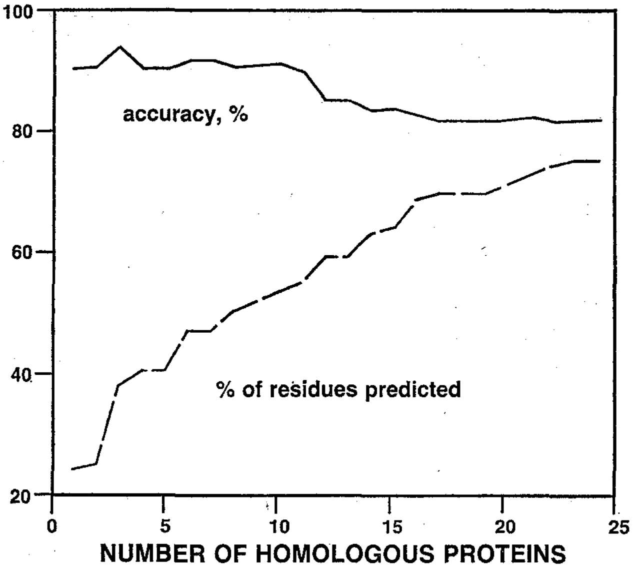

Fig. 2. The effect of homologous sequences on the accuracy of prediction (solid line) and the number of residues predicted (dashed line) of bovine pancreatic ribonuclease A. In the X-axis, the most homologous proteins were included in the prediction first (bovine pancreatic ribonuclease A itself was the first protein included) and then less homologous proteins were added as the value on the X-axis increases.

in the prediction (Table 1), the number of residues predicted was increased from 24 % to 52 % without losing reliability in the prediction (90–94 %). Further inclusion of fifteen less homologous proteins, in which 67–76 % of the residues were identical to bovine pancreatic ribonuclease A, increased the number of residues predicted up to 75 % of the whole sequence, although the reliability of the prediction was lowered to 82 %. This suggests that the inclusion of the proteins with less homology will increase the number of residues predicted but may affect reliability.

Since all of the five prediction methods used in this paper take account only the short- and medium-range interactions, our method is also limited to predict the secondary structures stabilized mainly by short- and medium-range interactions. If other factors such as long-range interactions, disulfide bond, cofactors, intermolecular association, environments other than water play a key role to stabilize the secondary structures, the reliability of the prediction may be low.

Recent DNA sequencing techniques are dramatically increasing the availability of homologous sequences which will make the combination of predictions of groups of homologous proteins an increasingly powerful tool to predict secondary structure for new protein sequences.

ACKNOWLEDGMENTS

We acknowledge K. Nagano, H.A. Scheraga and D.J. Lipman for providing their computer programs. We also acknowledge H. Nakashima for allowing us to use the protein sequence data. We thank T. Ooi, G.I. Glover and C.A. McWherter for useful discussions and suggestions.

REFERENCES

(1) Kabsch, W. and Sander, C. (1984) *Proc. Natl. Acad. Sci.,* USA **81,** 1075–1078.

(2) Nishikawa, K. and Ooi, T. (1986) *Biochim. Biophys. Acta.,* **871,** 45–54.

(3) Orcutt, B.C. and Dayhoff, M.O. (1982) Protein Sequence Database, National Biomedical Research Foundation, Washington, D.C.

(4) Garnier, J., Osguthorpe, D.J., and Robson, B. (1978) *J. Mol. Biol.,* **120,** 97–120.

(5) Chou, P.Y. and Fasman, G.D. (1978) *Adv. Enzymol.,* **47,** 45–148.

(6) Wako, H., Saito, N., and Scheraga, H.A. (1983) *J.Protein Chem.,* **2,** 221–249.

(7) Nagano, K. (1977) *J.Mol.Biol.,* **109,** 251–274.

(8) Nishikawa, K. (1983) *Biochim. Biophys. Acta.,* **748,** 285–299.

(9) Lipman, D.J. and Pearson, W.R. (1985) *Science,* **227,** 1435–1441.

(10) The 79 proteins denoted by the identification code of the Protein Data Bank are: 1AAT, 1ABP, 2ACT, 1ACX, 4ADH, 2ADK, 2APE, 2APP, 2ATC, 2B5C, 1BP2, 156B, 1C, 1CAC, 3CAT, 2-CDV, 2CHA, 3CNA, 5CPA, 1CPV, 1CRN, 1CTX, 3CYT, 2C2C, 1IC, 155C, 351C, 4DFR, 1ECD, 1EST, 3FAB (two chains), 1FC1, 2FDI, 3FXC, 3FXN, 1GCN, 2GPD, 2GRS, 1HIP, 1HMQ, 1INS (two chains), 4LDH, 1LHB, 1LH1, 2LYZ, 1LZN, 2MBN, 2MHB (two chains), 1MHR, 1MLT, 1NXB, 1OVO, 2PAB, 8PAP, 1PCY, 3PGK, 3PGM, 3PTP, 1PY, 1PPT, 4PTI, 1REI, 1RHD, 2RHE, 4RSA, 3RXN, 1SBT, 3SGB, 2SNS, 1SN3, 2SOD, 1SRX, 2SSI, 2TAA, 1TIM, and 3TLN.

(11) Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D,. Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J.Mol. Biol.* **112,** 535–542.

(12) Kabsch, W. and Sander, C. (1983) *Biopolymers,* **22,** 2577–2637.

(13) Warme, P.K., Momany, F.A., Rumball, S.V., Tuttle, R.W., and Scheraga H.A. (1974) *Biochemistry,* **13,** 768–782.