Prediction of protein structural similarities using a 3D-1D compatibility method

Yo Matsuo matsuo@peri.co.jp

Ken Nishikawa nishikawa@peri.co.jp

Protein Engineering Research Institute,
6-2-3 Furuedai, Suita, Osaka 565, Japan.

Abstract

The 3D-1D compatibility method is a new approach to protein structure prediction. It evaluates the compatibility of a one-dimensional (1D) amino acid sequence with known three-dimensional (3D) structures, and select the most likely structure. We have developed a method, which evaluates the 3D-1D compatibility using the following functions: side-chain packing, solvation, hydrogen-bonding, and local conformation functions. The method has been applied to a large number of sequences in databases. Here, the predictions of the structural similarities between the following pairs are described in detail: spermidine/putrescine-binding protein and maltose-binding protein, shikimate kinase and adenylate kinase, and mannose permease hydrophilic subunit ($IIAB^{Man}$) and galactose/glucose-binding protein. Functional and evolutionary implications of the predictions are discussed. Through these examples of predictions, the present work demonstrates the promise of the 3D-1D method.

1 Introduction

The genome projects are bringing us a wealth of DNA sequence data. We can extract a lot of biological information from the data. The information on the three-dimensional structure of a protein is very useful for understanding its biological function and its evolutionary relationship with other proteins. It is thus desirable to predict the structure of a protein from its amino acid sequence. However, the structure prediction has been a difficult task unless a sequence homology to a protein of known structure is found.

Recently, a major breakthrough in the structure prediction has been achieved. It is based on the observations of protein structures solved by X-ray crystallography and NMR spectroscopy.

松尾 洋，西川 建：蛋白工学研究所，〒565 吹田市古江台 6-2-3

Now we know a lot of examples of proteins which adopt similar structures despite little or no sequence similarity (e.g., actin and 70kD heat shock protein [1]). This means that the number of folds adopted by proteins is very limited. According to Chothia's estimate [2], it seems that there are only about 1,000 structural families coded in genomes. From these observations, a new approach to protein structure prediction, called the 3D-1D compatibility approach, has become the center of attention (see [3] for a review). The approach evaluates the compatibility of a one-dimensional (1D) amino acid sequence with known three-dimensional (3D) structures to see if the sequence adopts a structure similar to any of known structures.

We have developed a 3D-1D compatibility method [4, 5, 6], which uses side-chain packing, solvation, hydrogen-bonding, and local conformation functions. The method has been applied to a large number of protein sequences [5, 6, 7]. In the present paper, we describe the structure predictions of the following proteins: spermidine/putrescine-binding protein, shikimate kinase, and mannose permease hydrophilic subunit ($IIAB^{Man}$). Functional and evolutionary implications of the predictions are discussed. These examples of predictions demonstrate the usefulness of the 3D-1D method.

2 Materials and methods

Amino acid sequence data were derived from the NBRF-PIR sequence database release 38. The coordinate data were taken from the Protein Data Bank [8].

Four evaluation functions, side-chain packing (F_{sp}), solvation (F_{solv}), hydrogen-bonding (F_{hb}), and local structure (F_{loc}) functions, were used to evaluate the compatibility of an amino acid sequence with a structure. Except for F_{sp}, they were defined as in our previous work [4]. They have the following general form:

$$F_x = -\log(f_x(a;s)/f_x(s)), x = \{solv, hb, loc\},$$

where a denotes the type of amino acid residue (for F_{solv} and F_{loc}) or residue pair (for F_{hb}); s, the state of a (solvent-accessibility for F_{solv}, hydrogen-bonded or not for F_{hb}, and local structure for F_{loc}); $f_x(a;s)$, the frequency of a in the state s; and $f_x(s)$, the frequency of any residue or residue pair in the state s.

The side-chain packing function (F_{sp}) has been improved to take into account inter-residue contact and angle as well as distance [5, 6]. F_{sp} indicates the propensity of a residue pair (a,b) to be in contact in a particular spatial relationship. The spatial relationship between two residues was defined by the distance (d) between their Cβ atoms and the angle (θ) between the residues. The angle θ between residues i and j was defined as the sum of the angles Cβ_i-Cα_i-Cβ_j and Cβ_j-Cα_j-Cβ_i. Then, F_{sp} was defined by:

$$F_{sp}(a, b; d, \theta) = w(a, b; d, \theta)\{(dE_0(a, b) + dE(a, b; d, \theta)\}$$

Here,

$$w(a, b; d, \theta) = NC(a, b; d, \theta)/N(a, b; d, \theta),$$

$$dE_0 = -\log\{\frac{NC(a, b)/N10(a, b)}{NC/N10}\},$$

$$dE(a,b;d,\theta) = -\log\{\frac{NC(a,b;d,\theta)/NC(a,b)}{NC(d,\theta)/NC}\}.$$

$NC(a,b;d,\theta)$ is the number of observations of the residue pair (a,b) being in contact at a distance d and an angle θ; $N(a,b;d,\theta)$, that of (a,b) being at d and θ; $NC(a,b)$, that of (a,b) being in contact; $N10(a,b)$, that of (a,b) being within 10Åof each other; NC, that of any residue pair being in contact; $N10$, that of any residue pair being within 10Åfrom each other; $NC(d,\theta)$, that of any residue pair being in contact at distance d and angle θ. Now, $dE_0(a,b)$ describes the tendency of (a,b) to be in contact, and $dE(a,b;d,\theta)$ is the preference of (a,b) for being at d and θ, on the condition that they are in contact.

A sequence was threaded onto a structure using the Needleman-Wunsch algorithm [9] and the 3D-profile [10, 11]. For a sequence threaded onto a structure, scores S_x ($x = \{sp, solv, hb, loc\}$) were given by summing up the values of F_x over all residues or residue pairs. S_x scores were then added up to give S_{tot}, which measured the compatibility of the sequence with the structure. A more negative score indicates better compatibility.

A sequence was compared with a library of known structures. For the individual structures, compatibility scores S_{tot} were calculated. The scores were expressed in units of standard deviations from the mean (see [6] for details). It was empirically found that a score of -3.0 or better often indicates good compatibility.

3 Results and discussion

3.1 Spermidine/putrescine-binding protein

Spermidine/putrescine-binding protein (SPBP) is the periplasmic component of the spermidine/putrescine transport system of *E.coli* [12]. The SPBP sequence was compared with knwon structures. Then, maltose-binding protein (MBP; PDB code, 1OMP), which is another periplasmic binding protein, showed the best compatibility score (-3.66) (Table 1). This suggests that SPBP may adopt a similar structure.

The alignment of SPBP and MBP revealed a highly conserved sequence motif in the loop region between the first α-helix and the second β-strand. The conserved motif spans residues 53 to 61 ('FEKDTGIKV') of MBP and residues 46 to 54 ('FTKETGIKV') of SPBP. The search of the NBRF-PIR sequence database showed that iron-binding protein (IBP) also have the motif. From the sequence alignment, a consensus pattern of 'F(T/E)(K/R/Q)(D/E/A)TGIKV' was observed. The high specificity of the conserved motif to the three periplasmic binding proteins suggests a common functional role of the motif in the transport systems. By superimposing this region onto the MBP structure, we observe that the motif is located on the surface loop of the N-terminal domain (Figure 1). The motif might be involved in the interactions with the membrane components of the transport system.

3.2 Shikimate kinase

Shikimate kinase (SKase; EC 2.7.1.71) acts in the shikimate pathway for aromatic amino acid biosynthesis in plants and microorganisms, and catalyzes the phosphorylation of shikimic acid to shikimate 3-phosphate. The *E.coli* SKase II sequence was compared with known structures.

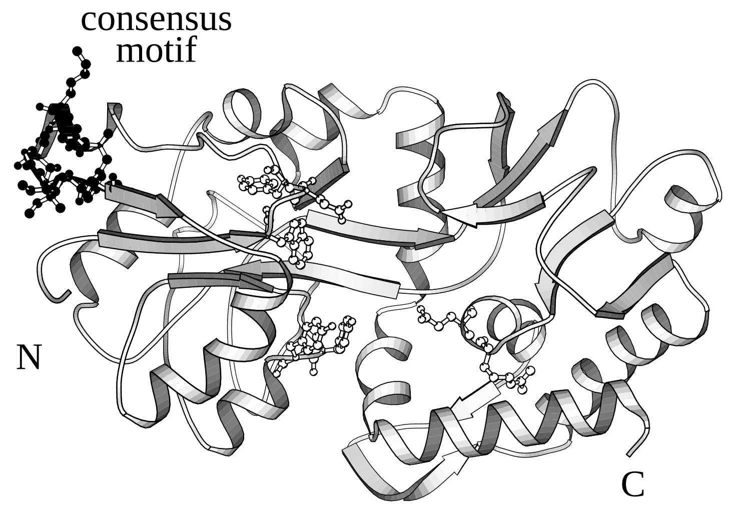

Figure 1: The SPBP sequence is threaded onto the maltose-binding protein structure. The black balls denote the residues of the consensus motif. The white balls denote the residues which might be involved in the substrate binding. The drawing was created using MOLSCRIPT [18].

rank	structure	PDB code	S_{tot}
1	maltose-binding protein	1OMP	-3.66
2	p-hydroxybenzoate hydroxylase	1PHH	-2.01
3	isocitrate dehydrogenase	3ICD	-1.97
4	sulfate-binding protein	1SBP	-1.87
5	actin	1ATNA	-1.83
6	ribose-binding protein	1DRI	-1.57
7	galactose/glucose-binding protein	2GBP	-1.55
8	phosphofructokinase	1PFKA	-1.53
9	malate dehydrogenase	4MDHA	-1.50
10	leucine-binding protein	2LBP	-1.50

Table 1: The compatibility of the SPBP sequence with known structures. The structures were sorted in order of their compatibility scores S_{tot}. The best 10 structures are listed.

The pig adenylate kinase (AKase; EC 2.7.4.3) structure (PDB code, 3ADK) showed the best compatibility score (-3.40), despite no significant overall sequence similarity (19% identity). The AKase structure also showed good compatibility scores for other shikimate kinase homologues. These results suggest the structural similarity between SKase and AKase.

The alignment of AKase and SKase shows that they have the type A sequence motif 'GXXXXGK(S/T)' [13]. The alignment also shows that two arginine residues of AKase, Arg107 and Arg138, which are involved in ATP-binding, are conserved among SKases. The conservation of these functionally important residues suggests the similarity between the functional mechanisms of AKase and SKase, and their evolutionary relationship as well.

3.3 Hydrophilic subunit of mannose permease

The mannose permease of *E.coli* is a component of the phosphotransferase system (PTS). It mediates the transport of mannose and related hexoses across the cytoplasmic membrane. The permease consists of a hydrophilic subunit $IIAB^{Man}$, and two transmembrane subunits IIC^{Man} and IID^{Man}. The hydrophilic subunit $IIAB^{Man}$, which is located in the cytoplasm, catalyzes the phosphate transfer from the histidine-containing phosphocarrier protein (HPr) to the sugar substrate.

The $IIAB^{Man}$ sequence was compared with known structures. The *E.coli* galactose/glucose-binding protein (GGBP) (PDB code, 2GBP) showed the best compatibility score (-3.34).

It has been reported that $IIAB^{Man}$ consists of two structural domains IIA (14kDa, residues 1-136) and IIB (20kDa, residues 156-323), which are linked by an Ala-Pro-rich flexible hinge of 20 residues [14]. The IIA and IIB domains aligned well with the N- and C-terminal domains of GGBP, respectively (Figure 2). The Ala-Pro-rich hinge aligned with the first α-helix of the C-terminal domain of GGBP. The helix is a part of the hinge region which is important for inter-domain motion of the periplasmic binding proteins.

$IIAB^{Man}$ is phosphorylated at His10 of the IIA domain and His175 of the IIB domain [14]. A phosphate group is transferred from HPr to His10, to His175, and to the sugar substrate.

rank	structure	PDB code	S_{tot}
1	adenylate kinase	3ADK	-3.40
2	leucine-binding protein	2LBP	-2.74
3	chloramphenicol acetyltransferase	3CLA	-1.70
4	glutathione peroxidase	1GP1B	-1.60
5	endothiapepsin	2ER7E	-1.60
6	xylose isomerase	6XIA	-1.44
7	arabinose-binding protein	8ABP	-1.36
8	p-hydroxybenzoate hydroxylase	1PHH	-1.36
9	β-lactamase	3BLM	-1.19
10	T4 lysozyme	3LZM	-1.18

Table 2: The compatibility of the *E.coli* shikimate kinase II sequence with known structures. The structures were sorted in order of their compatibility scores S_{tot}. The best 10 structures are listed.

rank	structure	PDB code	S_{tot}
1	galactose/glucose-binding protein	2GBP	-3.34
2	leucine-binding protein	2LBP	-2.53
3	malate dehydrogenase	4MDHA	-2.24
4	arabinose-binding protein	8ABP	-1.80
5	xylose isomerase	6XIA	-1.79
6	lactate dehydrogenase	6LDH	-1.35
7	tryptophan synthase β subunit	1WSYB	-1.34
8	aspartate aminotransferase	2AAT	-1.30
9	citrate synthase	2CTS	-1.27
10	aconitase	5ACN	-1.25

Table 3: The compatibility of the $IIAB^{Man}$ sequence with known structures. The structures were sorted in order of their compatibility scores S_{tot}. The best 10 structures are listed.

For this phosphotransfer to occur, the two histidine residues and the substrate binding region should be in close proximity. Our prediction satisfies this requirement; when the $IIAB^{Man}$ sequence is threaded onto the GGBP structure, His10 and His175 are located on loops in the substrate-binding cleft of GGBP (Figure 2).

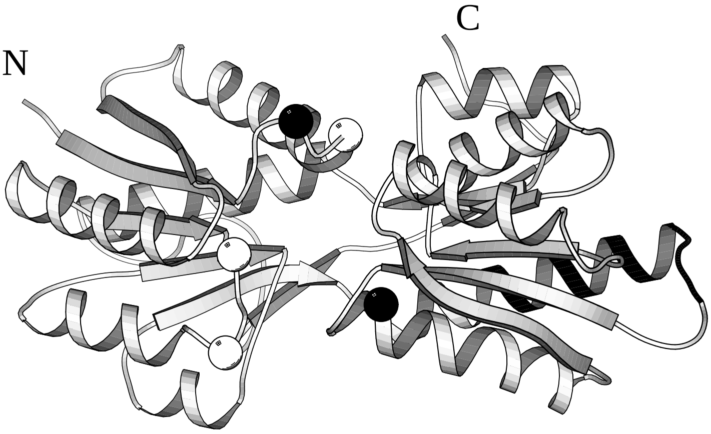

Figure 2: The $IIAB^{Man}$ sequence is threaded onto the GGBP structure. The black balls denote His10 (upper-left) and His175 (lower-right). The white balls denote the residues Trp12, Trp69, and Ser72 which are known to be in proximity of His10. The black ribbon denotes the Ala-Pro-rich hinge. The drawing was created using MOLSCRIPT [18].

Recently, Markovic-Housley et al. [15] predicted that the IIA domain would be structurally similar to flavodoxin, using 3D-profile methods [16, 17] and experimental data. Our results are consistent with their prediction. The N-terminal domain of GGBP, with which the IIA domain was aligned by our method, has the same topology with flavodoxin; the order of the five parallel β-strands is 54312.

Acknowledgements

We thank Shigeru Sugiyama and Kosuke Morikawa for discussions on the prediction of the SPBP structure, and Haruki Nakamura for the modeling of the SPBP structure.

References

[1] Flaherty, K.M., McKay, D.B., Kabsch, W. and Holmes, K.C. (1991) *Proc. Natl. Acad. Sci. USA*, 88, 5041-5045.

[2] Chothia, C. (1992) *Nature*, 357, 543-544.

[3] Wodak, S.J. and Rooman, M.J. (1993) *Curr. Opin. Struct. Biol.*, 3, 247-259.

[4] Nishikawa, K. and Matsuo, Y. (1993) *Protein Eng.*, 6, 811-820.

[5] Matsuo, Y. and Nishikawa, K. (1994) *FEBS Lett.*, 345, 23-26.

[6] Matsuo, Y. and Nishikawa, K. (1994) *Protein Sci.*, in press.

[7] Amano, T., Yoshida, M., Matsuo, Y. and Nishikawa, K. (1994) *FEBS Lett.*, 351, 1-5.

[8] 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.

[9] Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.*, 48, 443-453.

[10] Godzik, A., Kolinski, A. and Skolnick, J. (1992) *J. Mol. Biol.*, 227, 227-238.

[11] Wilmanns, M. and Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. USA*, 90, 1379-1383.

[12] Furuchi, T., Kashiwagi, K., Kobayashi, H. and Igarashi, K. (1991) *J. Biol. Chem.*, 266, 20928-20933.

[13] Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) *EMBO J.*, 1, 945-951.

[14] Erni, B., Zanolari, B., Graff, P. and Kocher, H.P. (1989) *J. Biol. Chem.*, 264, 18733-18741.

[15] Markovic-Housley, Z., Balbach, J., Stolz, B. and Génovésio-Taverne, J.-C. (1994) *FEBS Lett.*, 340, 202-206.

[16] Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) *Science*, 253, 164-170.

[17] Overington, J.P., Donnelly, D., Johnson, M.S., Sali, A. and Blundell, T.L. (1992). *Protein Sci.*, 1, 216-226.

[18] Kraulis, P.J. (1991) *J. Appl. Cryst.*, 24, 946-950.