Effect of Secondary Structure Prediction on Protein Fold Recognition and Database Search

Nickolai N. Alexandrov [1] nicka@amgen.com

Victor V. Solovyev[1] victors@amgen.com

[1] Amgen Inc. Thousand Oaks, California, 91320 U.S.A.

Abstract

Hydrophobic long-range interactions and local polypeptide chain propensities are the major factors directing protein folding. Incorporating both these terms in addition to the Dayhoff matrix helps us to increase quality of protein fold recognition via sequence-structure alignment. We have shown that the results of secondary structure prediction substantially increase a sensitivity of the fold recognition. To measure a performance of the protein fold recognition, we have developed a comprehensive test along with a set of the quality control scores based on the most populated structural families. With this test we have demonstrated improvement of the sequence alignment with consideration of the predicted secondary structure, even without knowledge of the real three-dimensional structure.

1 Introduction

The sequence similarity on the level of about 30% identities or higher implies a three-dimensional similarity (for long enough sequences). The opposite is not true: proteins from the same structural family do not always have similar sequences. It has been shown, however, that it is possible to find structurally reasonable alignment between two proteins even without significant sequential homology. This can be done when a 3D structure of one of the proteins is known. Then, based on the potentials, describing how well a giving residue from another protein can fit into an environment of the corresponding position in the three-dimensional structure of the first protein, an optimum sequence-structure alignment can be found (for review of different threading algorithms see Lemer et al., 1995). The potentials for sequence-structure compatibility may include secondary structure propensities of the different types of amino acids, pairwise residue-residue interaction, hydrophobic potentials, etc.

Sippl, 1990, suggested a self-threading test for evaluating the potential function threading sequences through all the structures without gaps. This test revealed that the most important terms in potential function are hydrophobicity and secondary structure propensity, which provide 90% of the successful recognition in the self-threading test (Alexandrov et al., 1996).

However, there was no good test for evaluating threading algorithms. In many cases, authors demonstrate a performance of their programs only on a few examples. Fischer et al., 1996, proposed a bigger test with 68 protein pairs and a quality score based on the top structure. Here we propose more comprehensive test for threading algorithms. With this test we can measure an effect of any modification in the potential function or aligning procedure.

Incorporation of the results of advanced secondary structure prediction method improves a quality of the fold recognition by 16% in compare with an ordinary sequence alignment. Further improvement has been achieved by adding contact capacity potentials and by imposing geometrical restrains on the alignment.

2 Test for threading

The goal of any threading algorithm is to recognize a structure which fits best to the query sequence. To evaluate the performance of the different threading procedures we have developed a test, based on the known protein structures. A list of 615 nonhomologous structures from the 35% PDB_SELECT database (Hobohm & Sander, 1994) was divided into structural families according to the SCOP classification (Murzin et al. 1995). Every protein sequence was thread through all the structures (except the native one for this sequence) from this list, producing a ranked list of the structures. In the case of ideal recognition, all the structures from the family of a test sequence should be on the top of the ranked list.

To measure the quality of recognition we used mainly two scores: N-score and S-score. N-score tells us what are the chances to get a correct fold on the first place, while S-score shows the average separation between proteins from the correct family and all others.

Let N_{pf} be a number of proteins in the family and N be a total number of protein structures. Threading sequence of protein i through all the structures results in the assigning to each structure j a compatibility score C_j^i. Usually, for a reasonable scoring function, $C_i^i = \max_j\{C_j^i\}$, i.e. a maximum value of the compatibility score is reached on the native structure. Here we do not consider native structures in our performance scores. To compute the performance score we sort the structures by the compatibility score C_j^i so that the first structure in the list has the largest compatibility score.

Let us define a delta-function δ_i^j as follows:

$$\delta_i^j = \begin{cases} 1, & \text{if } i \text{ and } j \text{ belong to the same family} \\ 0, & \text{if } i \text{ and } j \text{ belong to different families} \end{cases}$$

After threading all N_{pf} sequences, N-score is computed as a ratio of a number of the correct proteins on the first position to the total number of proteins in the family:

$$\text{N-score} = \frac{1}{N_{pf}} \sum_{i=1}^{N_{pf}} \delta_i^1.$$

In the case of ideal recognition N-score =1.

To take into account occurrences of family proteins at the beginning of the list, we use a more robust M-score:

$$\text{M-score} = \frac{1}{N_{pf}} \sum_{i=1}^{N_{pf}} \frac{1}{r_i},$$

where r_i is the smallest rank of a protein from the family.

We also use S-score to characterize a separation in the compatibility score between proteins from the family and all others:

$$\text{S-score} = \sum_{i=1}^{N_{pf}} \frac{< C_f^i > - < C_o^i >}{\frac{1}{2}(\sigma_f + \sigma_o)},$$

where $< C_f^i >$ and σ_f are the average value and standard deviation of the compatibility score for proteins from the family, $< C_o^i >$ and σ_o are the average value and standard deviation of the compatibility score for non-family proteins. S-score can be considered as an average Z-score for a protein family.

In this paper we computed the performance scores from the results, obtained on five structural families: immunoglobulins (27 proteins), globins (12, viral coat proteins (15), TIM-barrel structures (28), and four-helical cytokines (5).

First, we have checked the performance of the sequence alignment with Dayhoff matrix. We used Gotoh (Gotoh, 1982) algorithm to find an optimum alignment:

$$\begin{aligned}
E_{i,j} &= max\{D_{i,j-1} - a, E_{i,j-1} - b\} \\
F_{i,j} &= max\{D_{i-1,j} - a, F_{i-1,j} - b\} \\
D_{i,j} &= \max\{D_{i-1,j-1} + d_{i,j},\, E_{i,j}, F_{i,j}\} \qquad (1) \\
D_{0,0} &= E_{0,0} = F_{0,0} = 0; \\
E_{i,0} &= D_{i,0} = F_{0,i} = D_{0,i} = a + b(i-1).
\end{aligned}$$

The opening gap penalty $a = 10.0$, extension gap penalty $b = 1.0$. Values of d_{ij} equals to the Dayhoff matrix elements, corresponding to the residues on position i in structure and j in sequence. The results obtained for different families are shown in Figure 1.

A way to improve these results could be adding a secondary structure information to the scoring function.

An accurate prediction of secondary structure is exceptionally useful as a starting point for modeling of higher dimensional aspects of protein structure (Rost & Sander, 1995). The recent advances of fold recognition (or threading) technique (Fischer & Eisenberg, 1996; Russel et al, 1996) have been those where secondary structure predictions, and other protein characteristics were combined to suggest resemblance to an already known fold. The accuracy of secondary structure prediction methods has been improved significantly by the use of aligned protein sequences. The PHD method (Rost & Sander, 1993) and NNSSP method (Salamov & Solovyev, 1995) reach 71-72% of sustained overall three state accuracy when combine multiple sequence alignment with neural networks and nearest-neighbor algorithms, respectively. However, for

protein with no detected sequence homology, the best results, achieved by nearest-neighbor algorithms, was about 68% of sustained overall three-state (a-helix, b-strand and coil) accuracy (Yi & Lander, 1993; Salamov & Solovyev, 1995). In that case, Rost & Sander (1994) approach shows average accuracy of 63.1% only. Therefore in our investigation we use modified NNSSP approach to predict secondary structure of a query sequence.

3 Prediction of protein secondary structure by NNSSP method

The method is described in detail (Salamov & Solovyev, 1995) and we will only outline its main features and recent modifications.

The basic idea of the nearest-neighbor approach is the prediction of secondary structure state of the central residue of a test segment, based on the secondary structure of homologous segments from the proteins with known three-dimensional structure. The predicting type of secondary structure of a test residue was selected as the type of the majority of its nearest neighbors, i.e. as $\max(n_{\rm a},n_{\rm b},n_{\rm c})$, where $n_{\rm a}$,$n_{\rm b}$,$n_{\rm c}$ are the numbers of nearest neighbors with the helix, strand and coil types, respectively. The test residue was considered as the center of n consecutive amino acid residues of a sequence window. The nearest neighbors were selected by comparison the test window sequence against all n residue windows from the database using the similarity score measure (eqn. 2) averaged over all window residues. In the recent modification (Salamov, Solovyev, unpublished) the secondary structure of all position of a nearest neighbor was taken into account for prediction the states of aligned positions of the query sequence.

The key element in any nearest-neighbor prediction algorithms is a choice of a scoring table for evaluation of segments similarity. The local structural environment scoring developed by Eisenberg and coworkers (Bowie et al, 1991) assigns every residue of a protein with known three-dimensional structure to an "environment class" based on the local structural features of the residue position, such as the solvent accessibility, polarity and secondary structure. The score for matching a residue R_i with a local structural environment E_j was given by the informational statistics:

$$Score(R_i, E_j) = \log_{10}\left(\frac{P(R_i|E_j)}{P(R_i)}\right), \qquad (2)$$

where $P(R_i \mid E_j)$ is the probability of finding residue i in environment j, and $P(R_i)$ is the probability of finding residue i in any environment (Yi & Lander, 1993).

In NNSSP method additional environment classes as N- and C-ends of a-helices and b-strands have been created. Besides, b-turns separated from the other coil positions. In this way, 12 classes of secondary structure (5 for a-helices: internal, N- and C-caps, the left N- and the right C-adjacent positions; the analogous 5 for a b-strands, b-turns and coils) were combined with 6 categories of solvent accessibility/polarity, which give 72 environmental classes. 12 classes of secondary structure were used only for nearest neighbors selection (eqn. 2), but the only three-state secondary structure type (a, b or c) of the center residue of nearest neighbor windows was used for secondary structure assigning by majority rule. The score of matching a query residue with the dabase residue was computed as the environment score (2) plus a

score, estimated by mutation matrix. The total score for selection nearest-neighbor segments was calculated as an average score over windows of 17 amino acids.

The next improvement of the predicting accuracy was done by reducing the database where we search for amino acid fragments similar to a test protein sequence. We limited the database to a subset of proteins closest to the test protein in some general properties. Distance measure, based on the Chou-Fasman preference parameters (Chou & Fasman, 1978) for helices, strands and coils (D_{cf}) is

$$D_{cf} = \sum_{k=1}^{3} \left((1/l_t) \sum_{j=1}^{l_t} f_t^k(j) - (1/l_b) \sum_{j=1}^{l_b} f_b^k(j) \right)^2 , \qquad (3)$$

where $f_t(i)$ and $f_b(i)$ are frequencies of amino acid of type i; $f_t^k(j)$ and $f_b^k(j)$ are Chou-Fasman coefficients of the amino acid residue in the j-th position for the secondary structure type k (a, b, c); l_t and l_b are the lengths of a test and database proteins, respectively.

To exclude small elements, a simple filtering rules was applied: a) all helices of length 1 or 2 are converted to coils, except the case of bab which is converted to bbb; b) all strands of length 1 are converted to coils and c) all strands of length 2 surrounded by a-helical residues are converted to a-helices, i.e., abba to aaaa.

The modified NNSSP method provide 72% of sustained overall three state accuracy when we use multiple sequence alignment or about 69% accuracy for single sequence input, when tested on benchmark database of 126 nonhomologues proteins (Rost & Sander, 1994).

It is more informative to provide not only one state prediction for each position of a test sequence but compute a probability distribution P_i three possible states at each residue i. Let n_{a}, n_{b} and n_{c} are numbers of the best selected nearest neighbors for some positions of our database of proteins with known 3D structure. One can compute the proportion of a-, b- and c-states for these positions and use these data as probability estimation (Yi &Lander,1993). We scale our n_{a}, n_{b} and n_{c} values in 1 - 10 scale and produce $(10 \times 10 \times 10)$ matrix with probabilities belong to a definite state. These probabilities were used in scoring the resemblance to a tested fold providing better accuracy of recognition in comparing with one state secondary structure prediction.

4 Incorporating secondary structure into the potential function

With a knowledge of the real structure and predicted probabilities of the secondary structure, the term d_{ij} is replaced by the sum of the Dayhoff matrix (dm) and a contribution from a secondary structure prediction. Suppose the secondary structure of the i-th residue in the structure is s_i (s_i can be an alpha-helix, a beta-strand, or a coil), and the predicted probability of the j-th residue in sequence to be in the same conformation is $p(s_i, j)$. Then, in formula (1), $d_{ij} = dm_{ij} + 2 * (p(s_i, j) - 1/3)$. This modification results in significant improvement of the performance accuracy (Figure 1).

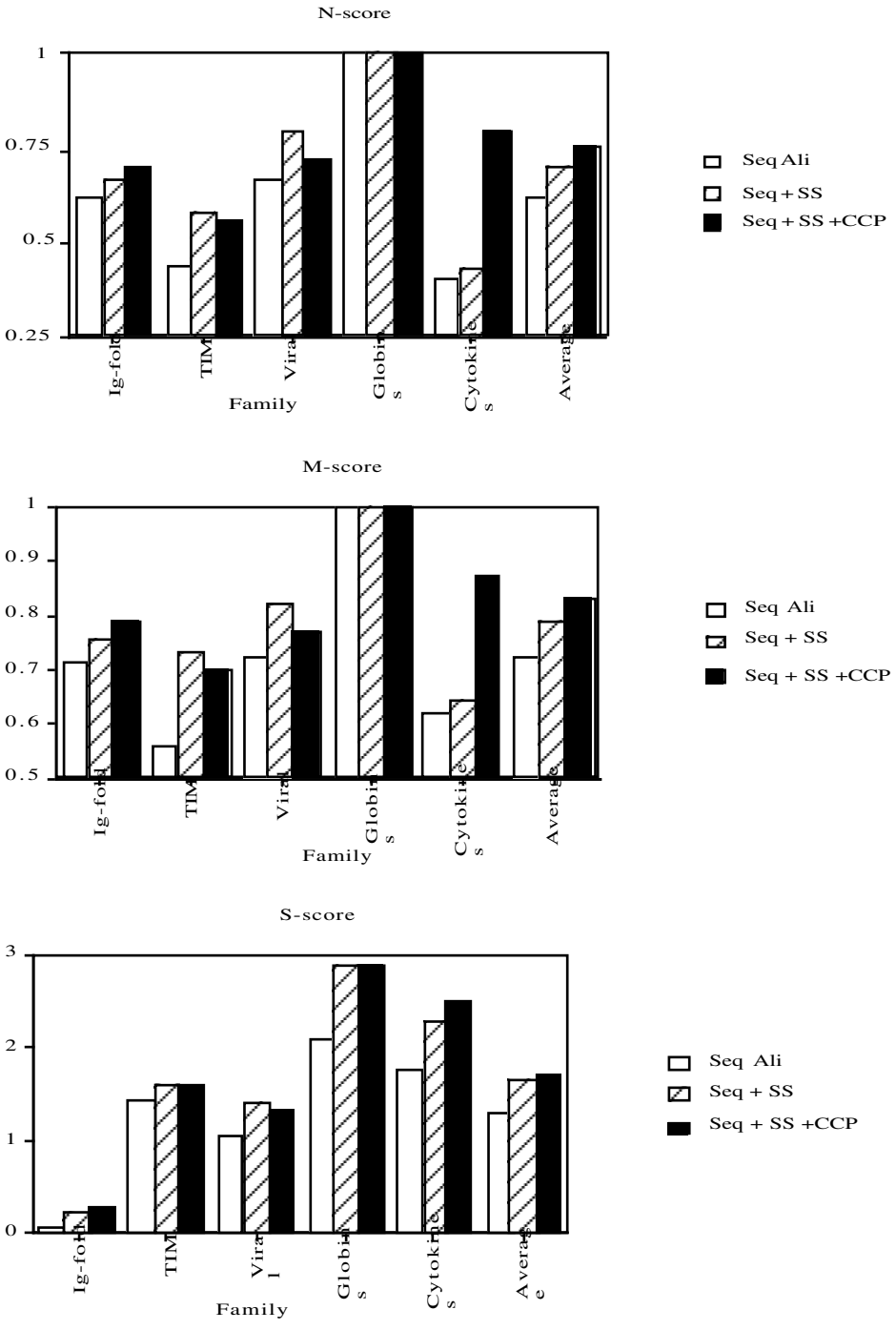

Figure 1: Quality of protein fold recognition for different families by pairwise sequence alignment, secondary structure prediction, and contact capacity potentials.

Table 1: Secondary structure prediction improves fold recognition even if 3D structure of proteins is not known.

Family	Seq Ali	Seq + SS
Immunoglobulin-like beta sandwich	0.11	0.15
beta/alpha (TIM)-barrel	1.43	1.63
Viral coat and capsid proteins	1.06	1.19
Globin-like	2.09	2.35
4-helical cytokines	1.77	1.79
Average	1.29	1.42

5 Contact capacity potentials

As described previously (Alexandrov et al. 1996), contact capacity potentials characterize an ability of different amino acids to make a certain number of contacts in protein structure. They reflect mainly hydrophobicity of residues. Adding this term to our potential function increases, in average, the performance scores (Figure 1).

6 Using secondary structure prediction for improvement of data base search

When 3D structures of query sequences and data base proteins are not known, we can predict their secondary structure and use this prediction in combination with mutation matrix for resemblance analysis. It can be expected that such procedure will permit us to recognize more distant homologous pairs than by simple sequence similarity. Secondary structure prediction methods reach good enough accuracy (about 70%) and the prediction errors occur mostly within short a- or b-structures, which are less responsible for protein fold configuration than long a-helices and b-strands. The methods use only sequence information, therefore any sequence database can be easily supplemented with predicted secondary structure. Using the NNSSP method we computed this information for our dataset of 5 protein families and for all other 615 PDB_SELECT proteins. Than we use the scoring system including sequence similarity and secondary structure to test this approach:

$$d_{ij} = dm_{ij} + 2*(\sum_{k=1}^{3}(2*p(k,j)+1/3)*(2*p(k,i)+1/3)-1/3).$$

7 Conclusion

We investigated the application of secondary structure prediction to the problem of fold recognition. Our work introduced several innovations to this approach. First, we developed new measures to evaluate the performance of recognition. Second, we constructed a scoring system

Table 2: Summary of a performance of different schemes for protein fold recognition.

Scheme	average S-score	average N-score	average M-score
Sequence alignment	1.29	0.63	0.72
Sequence alignment + secondary structure prediction	1.68	0.70	0.79
Sequence alignment + secondary structure prediction + contact capacity potentials	1.73	0.76	0.83

to combine mutation matrix score, secondary structure prediction (with probabilities) information and contact capacity potentials. Third, we incorporate secondary structure prediction into the database search when only sequence information is available. We have demonstrated a significantly better performance of sequence alignment and fold recognition with consideration of the predicted secondary structure (Table 2).

References

[1] lexandrov N.N., Nussinov R., Zimmer R.M. (1996). Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pacific symposium on biocomputing'96. eds. Hunter L. & Klein T.E., World Scientific, Singapore, pp.53-72.

[2] owie, J.U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164-170.

[3] hou P., & Fasman G. (1978). Empirical predictions of protein conformation. Annu. Rev. Biochem. 47: 251-276.

[4] ischer D. & Eisenberg D. (1996). Fold recognition using sequence-derived properties. Protein Sci. 5: 947-955.

[5] ischer D., Elofsson A., Rice D.W., Eisenberg D. (1996). Assessing the performance of invert protein folding methods by means of an extensive benchmark. Pacific symposium on biocomputing'96. eds. Hunter L. & Klein T.E., World Scientific, Singapore, pp. 300-318.

[6] otoh O. (1982). An improved algorithm for matching biological sequences. J. Mol. Biol. 162: 705-708.

[7] obohm U. & Sander C. (1994). Enlarged representative set of protein structures. Protein Sci. 3: 522.

[8] emer C.M.-R., Rooman M.J., & Wodak S.J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. Proteins 23: 337-355.

[9] urzin A.G., Brenner S.E., Hubbard T., & Chothia C. (1995). scop: a structural classification of protein database for the investigations of sequences and structures. J. Mol. Biol. 247: 536-540.

[10] ost B. & Sander C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Struc. Funct. Genet. 19: 55-72.

[11] ussel R.B., Copley R.R., & Barton G.J. (1996). Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 259:349-365.

[12] alamov A.A & Solovyev V.V (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247: 11-15.

[13] ippl M. (1990). Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213: 859-883.

[14] i T.-M. & Lander E.S. (1993). Protein secondary structure prediction using nearest-neighbor methods. J. Mol. Biol. 232: 1117-1129.