Reviving GOR method in protein secondary structure prediction: Effective usage of evolutionary information

Byung-Chul Lee[1],, Chang Jun Lee[2], Dongsup Kim[1*]

[1] Department of BioSystems, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[2] Department of Chemistry, Korea Advanced Institute of Science and Technology, Daejeon, Korea

*To whom correspondence should be addressed. E-mail: kds@kaist.ac.kr

Abstract

The prediction of protein secondary structure has been an important bioinformatics tool that is an essential component of the template-based protein tertiary structure prediction process. It has been known that the predicted secondary structure information improves both the fold recognition performance and the alignment accuracy. In this paper, we describe several novel ideas that may improve the prediction accuracy. The main idea is motivated by an observation that the protein's structural information, especially when it is combined with the evolutionary information, significantly improves the accuracy of the predicted tertiary structure. From the non-redundant set of protein structures, we derive the "potential" parameters for the protein secondary structure prediction that contains the structural information of proteins, by following the procedure similar to the way to derive the directional information table of GOR method. Those potential parameters are combined with the frequency matrices obtained by running PSI-BLAST to construct the feature vectors that are used to train the support vector machines (SVM) to build the secondary structure classifiers. Moreover, the problem of huge model file size, which is one of the known shortcomings of SVM, is partially overcome by reducing the size of training data by filtering out the redundancy not only at the protein level but also at the feature vector level. A preliminary result measured by the average three-state prediction accuracy is encouraging.

Introduction

The protein secondary structure prediction has been an essential bioinformatics tool. Thanks to the simplicity of problem description and the abundance of the data that can be used to develop a prediction method, the protein secondary structure prediction problem has been arguably the most popular bioinformatics research subject, exhausting almost all the imaginable algorithms [1]. The accuracies of the most accurate, state of the art, prediction methods are around 75% in terms of three-state prediction accuracy (Q_3) when they were applied to newly determined

protein structures [2]. One practical utility of the protein secondary structure prediction is that it greatly aids the protein tertiary structure prediction; it can significantly improve not only the fold recognition performance but also the alignment accuracy, resulting in producing more accurate three dimensional structure models [3]. There are three types of approaches: nearest neighbor methods such as NNSSP [4], statistical methods such as GOR [5], and machine learning approaches including the neural network (NN) [6] and the support vector machine (SVM) [7].

The key to one of the most successful prediction method, Psi-Pred [7], is the usage of the evolutionary information that can be obtained from the multiple sequence alignments. Following earlier work by Rost and Sander [8], D. Jones used the profiles of well-selected set of proteins as the input to his NN training. Before Rost and Sander's seminal work, the prediction accuracy (Q_3) was bellow 70% due to the absence of the evolutionary information in prediction algorithms. One of those methods, commonly known as GOR method [5], is based on statistical analysis on known protein structures and information theory. The prediction accuracy of the original GOR method is about 68%. Although its performance is worse than the popular "black box" machine learning approaches, it has one distinct advantage that all the parameters of GOR method have physical and statistical meaning, thus it can give insights into the relationship between sequence and structure. There was an attempt to combine GOR method and the evolutionary information to improve the accuracy of GOR method [9]. The authors of this paper utilized the fact that the proteins belonging to the same protein family tend to have similar structures. They were able to increase the accuracy of GOR method by predicting the secondary structures of not only a query protein but also its homologous proteins, and combining those predictions. A drawback of his approach is that the predictions should be made as many times as the number of the proteins homologous to a query protein. A better approach is that instead of using the directional information directly we first calculate the "homolog-averaged" directional information, which is the sum of directional information weighted by the amino acid frequencies observed at particular positions among the family members, and then use those parameters as an input to the machine learning algorithm such as SVM.

It is well known that the structural information, when it is used with the sequence information such as profile, significantly increases the performance of the template-based tertiary structure prediction [10]. The structural information is usually expressed as the knowledge-based potentials that are obtained by performing statistical analysis on known protein structures. Motivated by this observation, we first derive the "potential" for the secondary structure prediction by following the procedure similar to the way to derive the directional information table of GOR method, and then those potential parameters are combined with the frequency matrices obtained from the multiple sequence alignments to construct the feature vectors that are used to train the support vector machines (SVM) to build the secondary structure classifiers.

Methods

Structure Data

To derive "potential" parameters (will be described in the next subsection) for the secondary structure prediction similar to the directional information table of GOR, 1261 protein structures selected by PDB_SELECT [11] were used. They are non-redundant and of high resolution. We first calculated the eight-class secondary structures (H, G, I , E, B, C, T, and S) by running DSSP [12], and reduced these DSSP secondary structures to the three-class secondary structure types by following the CASP classification convention, i,e., Helix = (H, G, I), Beta strand = (E, B), and Coil (C, T, S). The training and test sets for the SVM training were prepared as follows; first, further reduce redundancy by keeping at most two proteins and removing the rest among the proteins with the same SCOP [13] family level classification, and then divide the remaining proteins into training set and test set in such a way that test set does not share any protein that has the same SCOP family level classification with any protein in training set, resulting in 339 proteins in training set and 339 proteins in test set.

Potential parameters

The "potential" parameters are obtained by estimating the log odds ratios,

$$I(S_i;R_j)=\log\frac{p(S_i,R_j)}{p(S_i)p(R_j)}, \quad \begin{matrix} i=1,2,...,n \\ j=i-7,...,i+7 \end{matrix}$$

where S_i is one of the three secondary structure type (H, E, C) at the position i, R_j the amino acid type at the position j, n the number of residues, $p(S_i)$ the probability of finding the secondary structure type S at the position i, $p(S_i, R_j)$ the probability of finding the amino acid type R at the position j, and $p(S_i, R_j)$ the corresponding joint probability.

Training support vector machines

Instead of using simply potential parameters, we calculated "homolog averaged" potential parameters;

$$\overline{I(S_i;j)}=\sum_{R=1}^{20} I(S_i;R_j)f(j,R),$$

where $f(j, R)$ is the frequencies of the amino acid type R at the position j that are found in the multiple sequence alignment by running the PSI-BLAST [14]. Because these "homolog averaged" potential parameters contain the evolutionary information that are the key to the protein structure, the feature vectors constructed from these parameters should contain more resolving power than simple potential parameters. It should be recognized that by doing so we effectively combine GOR method and the evolutionary information similar to the previous work [9]. Moreover, our method is superior to the previous work [9] because we do not need to repeat the same prediction for all the protein family members. The final feature vectors comprise of the frequencies of occurrences of each amino acids and the homolog-averaged potential parameters for three secondary structure types

within window size of 11. Therefore, the dimension of the feature vectors is 253(=11*20+11*3).

The publicly available SVM^{light} [15] was used to train SVMs. To reduce the number of feature vectors, while maintaining the performance, we removed about half of feature vectors by calculating Euclidean distances between all the feature vectors, and removing about half of feature vectors in such a way that none of all possible pairs of feature vectors in the final training data have smaller Euclidean distance than the specified cutoff value. We construct three binary classifiers, H/~H, E/~E, and C/~C, where ~ denotes negation. The radial basis functions (rbf) with length-scale parameter g,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-g\|\mathbf{x}_i - \mathbf{x}_j\|^2),$$

were used for the kernel functions. To handle unbalanced data, we employed the technique that used the different penalty parameters for true and false data. As suggested by Platt [16], the outputs from the SVM $y(\mathbf{x})$ for the feature vectors $\mathbf{x}$ are transformed to the posterior probabilities,

$$P(y \mid \mathbf{x}) = \frac{1}{1+\exp[-ay(\mathbf{x})+b]}$$

where a and b are the adjustable parameters that were chosen to have the optimal performance. To make prediction, the secondary structure type with the maximum posterior probability $P(y|\mathbf{x})$ was selected, and the three-state prediction accuracy Q_3 defined by,

$$Q_3 = \frac{\#\text{ of residues correctly predicted}}{\#\text{ of prediction made}} \times 100,$$

was calculated. Also for each state I=(H,E,C), Q_I, defined by the % fraction of the number of residues *correctly predicted* in state I out of the number of residues observed in state I, was calculated.

Results and Discussion

We have tried several kernel functions including linear, polynomial and rbf. As show in Table I, the rdf with the length-scale parameter g = 0.1 seemed the optimal choice. For the regularization parameter, we have not tried to optimize it, but rather simply used the default values set by the program SVM^{light}. It is also easy to recognize from Table I that too large g values tend to over-train the classifiers.

Table I The training and testing accuracies (%) of three binary classifiers with various choice of the parameter g.

Classifier	g = 0.05	g = 0.1	g = 0.2
H/~H			
training	86.14	87.90	89.51
testing	84.24	84.44	83.52
E/~E			
training	88.72	87.90	90.96
testing	87.16	87.38	86.64
C/~C			
training	79.23	82.34	86.84
testing	76.54	76.67	75.95

A well-known problem of SVM is that the fraction of training examples that become support vectors is rather high, as shown in Table II, which requires huge memory and long computation time for prediction. Therefore, it is crucial to reduce the number of training examples while

maintaining the accuracy of classifiers. We were able to achieve this objective by filtering procedure that was explained in the previous section without hurting the performance of the classifiers.

Table II The percentage of the training examples that become support vectors for three classifiers

Classifier	%SVs
H/~H	45.50
E/~E	36.48
C/~C	59.98

Although the testing sets are different, the classification accuracies of current work seem better than those of previous work by Hua and Sun (HS) [7], as shown in Table III.

Table III The accuracies (%) of binary classifiers. HS/RS126 and HS/CB513 refer to the results on RS126 and CB513 sets, respectively, by Hua and Sun [7].

Classifier	HS/RS126	HS/CB513	Present
H/~H	80.36	83.02	84.44
E/~E	81.25	83.39	87.38
C/~C	73.20	75.52	76.67

It should be pointed out, however, that it is premature to argue that current method is better than that of HS because the testing sets are different. Nonetheless, it is worth mentioning that the accuracies of current work are higher than those for the benchmark set CB513 that has been know to be an easy benchmark set.

Finally, the average three-state prediction accuracy (Q_3) of our method is shown and compared with those by HS in Table IV.

Table IV Prediction accuracies (%) of present method compared with those of HS/RS126 and HS/CB513. For description of HS/RS126 and HS/CB513, see the caption of Table III.

Method	Q_3	Q_H	Q_E	Q_C
Present	73.31	76.80	56.41	79.49
HS/RS126	71.1	72.0	56.1	77.2
HS/CB513	72.9	74.8	58.6	79.0

Although it is not directly comparable, it is reasonable to say that the performance of our method compares favorably against that by HS.

The prediction accuracy of present method at current stage of development is roughly 2% lower than those of the most accurate secondary structure prediction programs. There are several reasons for lower performance. First, the number of proteins in training set is relatively small, compared with that of Psi-Pred [6], which has thousands of proteins as training purpose. It is planned to add more proteins in our training set, up to the level of Psi-Pred. Second, many parameters are not optimized yet. The only parameter that we have tried to optimize is the parameter g. It is likely that the optimization of the regularization parameter is important for the performance enhancement. Third, we have only tried one multi-class classification scheme. It is clear that we need to develop more elaborate multi-class classification scheme. Despite all these, the performance of our program is reasonably good at this stage of development. Therefore, we would like to mention that our preliminary result is highly encouraging.

Acknowledgement

This work is supported by CHUNG Moon Soul center for BioInformation and BioElectronics (CMSC).

References

[1] B. Rost, Protein structure prediction continues to rise. *J. Struct. Biol.* 134:204-218 (2001).

[2] I. Y. Y Koh, et al. EVA: evaluation of protein structure prediction servers. *Nucl. Acids. Res.* 31:3311-3315 (2003).

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

[4] A. A. Salamov and V. V. Solovyev, Prediction of protein secondary structure by combining nearest-neighbor algorithm and multiple sequence alignments. *J. Mol. Biol.* 247:11-15 (1995).

[5] J. Garnier, J. F. Gibrat, and B. Robson, GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266:540–553 (1996).

[6] D. Jones, Protein secondary structure prediction based on position specific scoring matrices. *J. Mol. Biol.* 292:195–202 (1999).

[7] S. Hua and Z. Sun, A novel method of protein secondary prediction with high segment overlap measure: support vector machine approach, *J. Mol. Biol.* 308:397-407 (2001).

[8] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–599 (1993).

[9] A. Kloczkowski, K.-L. Ting, R.L. Jernigan, and J. Garnier, Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence, *Proteins*, 49:154-166 (2002).

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

[11] U. Hobohm and C. Sander, Enlarged representative set of protein structures, *Protein Sci.* 3:522-524 (1994).

[12] W. Kabsh and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded geometrical features. *Biopolymer*, 22:2577-2637 (1983).

[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247:536-540 (1995).

[14] S. F. Altschul, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25:3389–3402 (1997).

[15] T. Joachims, Making large-scale SVM learning practical. In B. Scholkopf et al. eds. *Advances in kernel methods-support vector learning*. MIT Press, Cambridge, MA.

[16] J. C. Platt, Probabilistic outputs for SVMs and comparisons to regularized likelihood methods, In A. Smola et al. eds. *Advances in large margin classifiers*. MIT Press, Cambridge, MA.