Original Paper

Prediction of Protein-Protein Interaction Sites Using Only Sequence Information and Using Both Sequence and Structural Information

Masanori Kakuta,[†1] Shugo Nakamura[†1] and Kentaro Shimizu[†1]

Protein-protein interactions play an important role in a number of biological activities. We developed two methods of predicting protein-protein interaction site residues. One method uses only sequence information and the other method uses both sequence and structural information. We used support vector machine (SVM) with a position specific scoring matrix (PSSM) as sequence information and accessible surface area (ASA) of polar and non-polar atoms as structural information. SVM is used in two stages. In the first stage, an interaction residue is predicted by taking PSSMs of sequentially neighboring residues or taking PSSMs and ASAs of spatially neighboring residues as features. The second stage acts as a filter to refine the prediction results. The recall and precision of the predictor using both sequence and structural information are 73.6% and 50.5%, respectively. We found that using PSSM instead of frequency of amino acid appearance was the main factor of improvement of our methods.

1. Introduction

Protein-protein interactions play an important role in a number of biological activities such as DNA replication and repair, molecular recognition, enzyme reaction, and signal transduction cascade. The yeast two-hybrid system was recently developed as high-throughput method of screening protein-protein interactions[1),2)]. However, to study protein function in detail and to design drugs, it is important to know how protein-protein interactions occur at the atomic level. Many X-ray diffraction and nuclear magnetic resonance (NMR) experiments have been conducted to obtain detailed structural information of protein complexes. However, these experiments are time consuming and expensive. To resolve these problems, various computational methods have been developed.

Protein-docking methods, which predict structures of a complex based on the structures of their protomers, are common examples of such computational methods[3)]. Various docking methods have been evaluated in a community-wide experiment, Critical Assessment of PRediction of Interactions (CAPRI)[4),5)]. In these methods, protein-protein interaction sites determined by experiment or predicted by computation reduce required search space and filter out incorrect models[4)–6)]. Prediction of protein-protein interaction sites is useful for docking methods, mutational experiments, and drug design. Many methods of predicting protein-protein interaction sites are also useful for determining the candidate sites for docking. In addition, their application might be broader than docking in that they require no partner information.

Since knowledge of the characteristics of protein-protein interaction sites is useful for detecting such sites, many studies have analyzed various interface properties such as physicochemical properties[7)], residue propensities[8)], and evolutionary conservation[9)]. Various prediction methods using the properties of protein-protein interaction sites described above have been developed. Jones and Thornton used six parameters described above to predict whether a surface patch is an interface or not[10)]. ProMate used various physicochemical properties in addition to amino acid pairing preferences and evolutionary conservation[11)]. The optimal docking area method utilized only desolvation energy and found that the low energy point was located in the known binding site or in its vicinity[12)]. The evolutionary trace method[13)–15)] and other methods of calculating the conservation score[16),17)] used multiple sequence alignment to predict functionally impor-

[†1] Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo

tant residues such as the protein-protein interaction site, the protein-DNA binding site, and the enzymatic active site.

In recent years, machine learning algorithms such as artificial neural network (ANN) and support vector machine (SVM) have been used for prediction. Machine learning methods have few parameters to determine and are suitable for combining various features. There are methods that use only sequence information[18),19)] and that use both sequence and structural information[19)–23)]. Prediction using structural information is subdivided into two categories: interaction residue prediction[19)–22)] and interaction patch prediction[23)]. Interaction residue prediction predicts whether a given residue is in the interaction site or not. Such information is not always necessary because sometimes researchers only need to know whether the residue is in the vicinity of an interaction site. In predicting interaction patches, however, results depend on the definition of patch and the residue-base prediction realizes more fine-grained prediction than does the patch prediction. We developed our interaction residue prediction methods with these considerations in mind.

Here, we present two methods of predicting protein-protein interaction site residues: a method using only sequence information and a method using both sequence and structural information. We used SVM as a classifier to decide each residue as an interaction site residue or a non-interaction site residue. The effectiveness of SVM is well known in various predictions such as secondary structure[24)], accessible surface area[25)], and fold recognition[26)]. The new and original points of our study are the features input to SVM and the application of SVM in two stages. We use the position specific scoring matrix (PSSM) of PSI-BLAST as a sequence feature and accessible surface area (ASA) of polar and non-polar atoms in a residue as a structural feature. These features for sequentially or spatially neighboring residues were combined to constitute a feature vector. We show that ASAs of polar and non-polar atoms in a residue is superior to relative ASA of a residue, which is normalized by its maximum value, in predicting protein-protein interaction residues and that it is important to use PSSM instead of frequency. In addition, we use SVM in two stages to filter out isolated predictions, which are predictions contrary to those of surrounding residues.

2. Materials and Methods

2.1 Datasets

A dataset consisting of 563 nonhomologous protein chains with no more than 25% sequence identity, which is the same set as used by Koike and Takagi[19)], was used to evaluate our method's performance. This dataset consists of protein chains in the Protein Data Bank (PDB)[27)] that satisfy the following conditions: 1) those are determined by X-ray with resolution better than 3.5 angstroms, 2) the distance between the nearest heavy atoms in other chains is within 5 angstroms, and 3) the length of their chains is longer than 100 residues. To exclude complexes that may not form complexes in vivo, we omitted hetero complexes with <20 interfacial residues and homo complexes with <30 interfacial residues. Complexes whose BLAST[28)] E-value is larger than 0.01 are defined as hetero complexes. All other complexes are defined as homo complexes. A dataset consisting of 271 hetero complex chains and 292 homo complex chains was obtained with these definitions.

2.2 Definition of Protein Interaction Site Residues

The solvent accessible surface area of each residue was computed with the Dictionary of Protein Secondary Structure (DSSP) program[29)]. Residues with surface areas more than 10% exposed to solvent were defined as surface residues. Interaction site residues were defined as surface residues where the distance between any heavy atoms in residue and any heavy atoms in the interacting proteins was within 5 angstroms. About 23% of whole residues (155054 residues) and 30% of surface residues (104331 residues) in our dataset were interaction site residues (31816 residues) under this definition. In this study, we predicted interaction site residues from whole residues when sequence information alone was used and from surface residues when structural information was also used.

2.3 Outline of Prediction

Our method uses SVM in two stages to filter out isolated predictions, which are predictions contrary to those of surrounding residues (**Fig. 1**). In the first stage, only the sequence or both sequence and structural features were extracted. These features of sequentially or spatially neighboring residues were combined

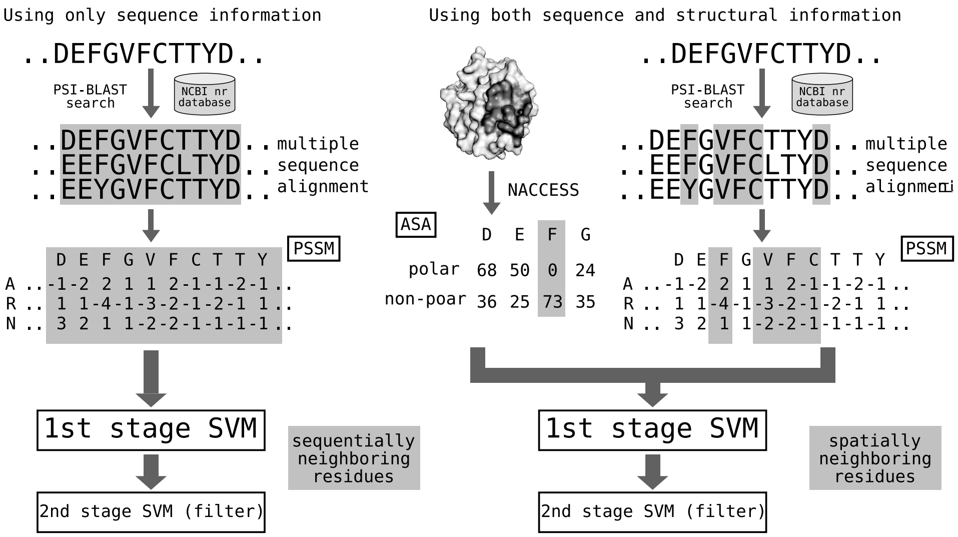

Fig. 1 Outline of our prediction methods. The left figure represents the method using only sequence information and the right figure represents the method using both sequence and structural information.

to constitute a feature vector for first stage prediction. The SVM prediction produced a decision value for each residue. In the second stage, the decision values of neighboring residues were combined in the same way to constitute a feature vector for second stage prediction. Five-fold cross-validation was used to estimate the performance

2.4 Feature Extraction

Frequencies [19),20),22)] and PSSMs [21)] have been used as sequence information for machine learning approaches. However, which is better in predicting interaction residues has not yet been studied. Thus, we compared the prediction accuracy of frequencies and PSSMs. Frequency is an N-by-20 matrix for sequences of length N. The (i,j) element of this matrix represents the ratio of amino acid type j at the i-th sequence position. Position specific scoring matrix (PSSM) of PSI-BLAST,[28)] is also an N-by-20 matrix for sequences of length N. The (i,j) element of this matrix represents the degree of conservation of amino acid type j at i-th sequence position. To make frequencies and PSSMs for each protein chain, PSI-BLAST with two iterations against NCBI nr database was used. All PSI-BLAST arguments except for iteration were default values. Both frequency and PSSM, which are created from multiple sequence alignment, generate a 20-dimentional vector for each residue of each protein. However, PSSM takes both sequence weighting and pseudocount frequencies into consideration. Frequency is simply the ratio of the amino acid types in each position. This means that PSSM represents the similarity between feature vectors more accurately than frequency.

Relative ASA has been used to represent exposure of residue to solvent and was used to predict interaction residues [19)]. The hydrophobic residues, which have a large relative ASA, are likely to be interaction sites because such areas are stabilized by interaction with hydrophobic regions of other protein chains. Since we expected that looking at the exposure of a hydrophobic area at the atomic level rather than the residue level improves prediction performance, we subdivided the ASA of a residue into the ASAs of polar and of non-polar atoms in a residue. ASAs of polar and non-polar atoms in a residue represent the region of residue that is exposed to solvent. ASAs of polar and non-polar atoms for each residue were calculated by the NACCESS program [30)]. Polar atoms are all oxygen and nitrogen atoms, and non-polar atoms are the others.

2.5 Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm for two-group classification problems [31)]. SVM is known for its high performance in classifying unknown data and has been applied to many problem areas [24)–26)]. SVM maps the feature vector into a high dimensional feature space and classifies the samples by separating the hyperplane in this space. At the training stage, SVM searches for an optimal hyperplane by solving a quadratic programming optimization problem. This hyperplane, determined by the criterion that maximizes the distance of nearest feature vector, has good generalization performance. We used LIBSVM, Library for Support Vector Machines[32], with a radial basis function (RBF) kernel to predict protein-protein interaction site residues. The SVM using the RBF kernel has two parameters, gamma and cost. Gamma determines RBF kernel function. Cost determines softness of the hyperplane. We fixed gamma at a default value of LIBSVM (1/dimension of feature vector) and tried 1, 10, 20, ..., 90, and 100 as cost values to search for the best parameter set.

2.6 SVM Training and Prediction

In the method using only sequence information, PSSMs of 11 sequentiall neighboring residues were used as features for the training and the prediction of the first stage SVM. This vector was used to predict whether the central residue is an interaction residue or not. In the method using both sequence and structural information, PSSMs and ASAs of the central residue and 14 spatially neighboring residues were used as features. ASAs of polar and non-polar atoms in a residue were treated separately. Spatially neighboring residues were defined as the nearest residues as measured by distance between the alpha carbon of the central residue and that of another residue. These features were sorted in ascending order of distance when we create a feature vector. The value of each element of a feature vector is scaled to the range [0, 1] because the range of the PSSM value and that of ASA are different. The feature vectors were used in prediction, and the decision values, which are calculated by the decision function, were obtained for each residue. Usually a decision value is digitized by the sign function, and the value indicates each class. However, in this approach, raw decision values of 11 sequentially or 15 spatially neighboring residues were used as features for SVM input in the second stage. The output of the second stage is also raw decision values. We predict protein-protein interaction site residues based on these values and adjust recall-precision performance by adjusting cutoff values (default value is 0) for comparison with other methods. If the decision value is higher than or equal to the cutoff value, the residue is predicted to be an interaction site residue and if the decision value is lower than the cutoff value, the residue is predicted to be a non-interaction site residue. In the method using both sequence and structural information, the feature vector has 330 dimensions (20 of PSSM plus 1 of ASA of polar atoms plus 1 of ASA of non-polar atoms per residue for 15 residues) in the first stage and has 15 dimensions (one of decision value per residue for 15 residues) in the second stage. The performance of each predictor was evaluated by 5-fold cross-validation.

2.7 Measure of Prediction Performance

Because of disproportion between the number of interaction site residues and of non-interaction site residues, evaluation of performance based only on accuracy is inadequate. Thus we considered the following to evaluate predictor performance.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$
$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{\mathrm{AP} \times \mathrm{AN} \times \mathrm{PP} \times \mathrm{PN}}}$$

TP, TN, FP, FN, AP, AN, PP, and PN mean the numbers of true positives, true negatives, false positives, false negatives, all positives, all negatives, predicted positives, and predicted negatives, respectively. Accuracy is the percentage of correct predictions in all predictions. Recall is the percentage of correctly predicted interaction site residues in interaction site residues. Precision is the percentage of correctly predicted interaction site residues in predicted interaction site residues. Matthew's correlation coefficient, MCC, represents how well a predicted class correlates with an actual class and ranges from -1 to 1. MCCs of 1, 0, and -1 mean perfect correlation, decorrelation, and inverse correlation, respectively.

3. Results

3.1 Prediction Performance

For prediction using both sequence and structural information, the recall and precision of the method using two-stage SVM with PSSMs and ASAs of polar and non-polar atoms as fea-

Table 1 Performance of predictor using 15 spatially neighboring residues.

SVM Predictor	Accuracy (%)	Precision (%)	Recall (%)	MCC
2-stage SVM using PSSMs and aASAs	69.4 (51.5)	50.0 (30.4)	75.5 (46.1)	0.391
1-stage SVM using PSSMs and aASAs	69.4 (52.0)	50.0 (30.4)	73.6 (44.9)	0.382
2-stage SVM using PSSMs and rASAs	69.6 (52.0)	50.1 (30.4)	73.6 (44.8)	0.384
1-stage SVM using PSSMs and rASAs	69.6 (52.5)	50.1 (30.4)	71.4 (43.4)	0.374
2-stage SVM using frequencies and aASAs	69.3 (52.5)	49.9 (30.4)	71.1 (43.5)	0.369
1-stage SVM using frequencies and aASAs	69.4 (53.1)	50.0 (30.4)	68.7 (41.9)	0.359
2-stage SVM using frequencies and rASAs	69.5 (53.0)	50.0 (30.4)	69.2 (42.2)	0.363
1-stage SVM using frequencies and rASAs	69.6 (53.7)	50.2 (30.4)	66.2 (40.3)	0.351

Randomly predicted values are shown in parentheses. Each random value was calculated as described below. $Random\_accuracy = (1 - Random\_precision) \times (1 - Random\_recall) + Random\_precision \times Random\_recall$. $Random\_precision = (TP + FN) \div (TP + TN + FP + FN)$. $Random\_recall = (TP + FP) \div (TP + TN + FP + FN)$

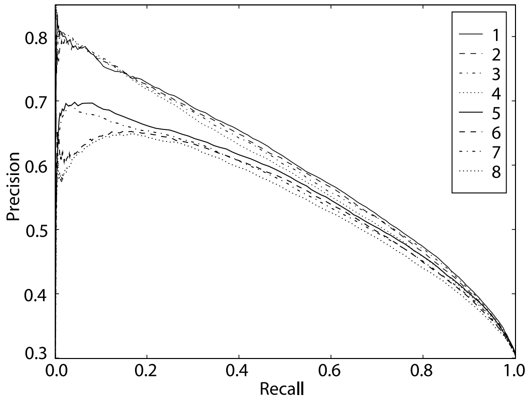

Fig. 2 Recall-precision curves of eight predictors using both sequence and structural information.
1: 2-stage SVM using PSSMs and aASAs.
2: 1-stage SVM using PSSMs and aASAs.
3: 2-stage SVM using PSSMs and rASAs.
4: 1-stage SVM using PSSMs and rASAs.
5: 2-stage SVM using frequencies and aASAs.
6: 1-stage SVM using frequencies and aASAs.
7: 2-stage SVM using frequencies and rASAs.
8: 1-stage SVM using frequencies and rASAs.
aASA means ASA of polar and non-polar atoms in a residue. rASA means relative ASA of a residue.

ture vectors were 73.6% and 50.5%, respectively, of default cutoff value. To compare the performance of different methods, we examined the performance of seven other methods. **Table 1** and **Fig. 2** show those performance results. Since the precisions of the eight methods ranged between 47.6% and 50.9% of default cutoff value, the precision in Table 1 was adjusted to about 50% for comparison by adjusting cutoff values. All three techniques (PSSMs, ASAs of polar and non-polar atoms, and two-stage SVM) contribute to improvement of performance in all ranges, and the rate of improvement in performance is higher in higher precision regions.

The contribution of PSSMs is especially important because the performances of the methods using the PSSMs are higher than the performances of the methods using frequencies. The PSSMs incorporate sequence weighting and pseudocount frequencies whereas the frequencies are merely the ratio of amino acids on each position. The PSSM is useful for extracting the similar regions of multiple sequence alignments. In this study,the PSSM was constructed by PSI-BLAST with two iterations. We made the profiles with more than two interactions but the prediction accuracies were not improved. For more iterations, distantly related proteins can be incorporated in PSSMs but false positives may also be increased.

The results for prediction using sequence information alone are shown in **Table 2** and **Fig. 3**. The protein-protein interaction site residues were predicted using PSSMs and frequencies of 11 sequentially neighboring residues. The performances of the methods using only sequence information are generally worse than that of the method using both sequence and structural information. This indicates that structural information is important in predicting protein-protein interaction sites. The method using PSSMs performed better than the one using frequencies. However, using two-stage SVM did not improve performance. There are two possible reasons for this. One is that the tendency of sequentially neighboring residues to form clusters of interaction residue is weaker than for spatially neighboring residues. This weak tendency may not be sufficient to enable reprediction. The other is that, in first stage prediction, the method using sequence information alone does not perform as well as the method using structural information. Reprediction based on such inaccurate prediction is meaningless.

The performance of the method using

Table 2 Performances of predictors using 11 sequentially neighboring residues.

SVM Predictor	Accuracy (%)	Precision (%)	Recall (%)	MCC
1-stage SVM using PSSM	57.9 (51.2)	30.0 (22.9)	62.3 (47.6)	0.160
2-stage SVM using PSSM	57.9 (51.1)	30.0 (22.9)	62.9 (48.0)	0.162
1-stage SVM using frequencies	60.8 (55.0)	30.0 (22.9)	53.2 (40.7)	0.139
2-stage SVM using frequencies	60.4 (54.6)	30.0 (22.9)	54.2 (41.4)	0.141

Randomly predicted values are shown in parentheses. Since precision of two methods were 30.0% and 30.8% in default cutoff value, precision was adjusted to 30.0% for comparison by adjustment.

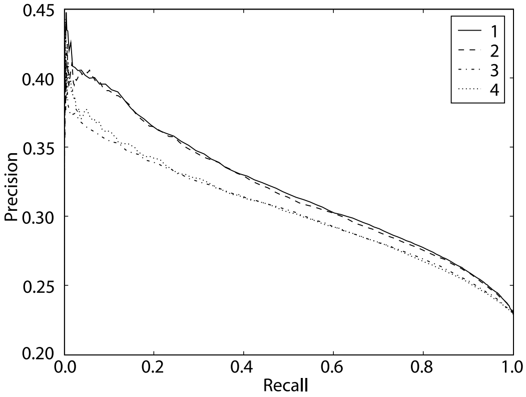

Fig. 3 Recall-precision curves of four predictors using only sequence information. **1**: 1-stage SVM using PSSMs. **2**: 2-stage SVM using PSSMs. **3**: 1-stage SVM using frequencies. **4**: 2-stage SVM using frequencies.

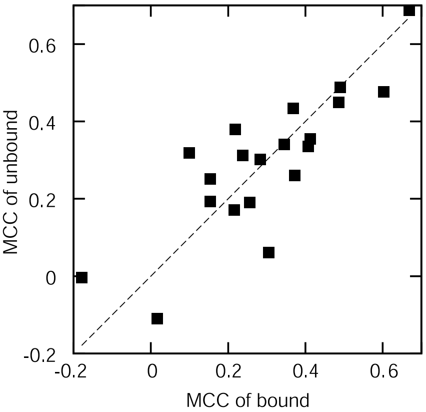

Fig. 4 MCCs of unbound structures were plotted against MCCs of bound structures.

two-stage SVM with PSSMs and ASAs of polar and non-polar atoms in a residue was compared to Koike and Takagi's method[19], which used the same dataset for training and prediction. Their method used single-stage SVM with frequencies and relative ASA of residue. The recall and precision of their method were 44.6% and 56.1%, respectively. We obtained the result that the recall was 73.6% (29% higher than theirs) when the precision was 50.5% (5.6% lower than theirs). To directly compare two methods, we also adjusted the threshold of our method. The recall of our method was 62.2% when the precision was 56.1%. The precision was 63.6% when the recall was 44.6%. Thus, the performance of our method was higher than theirs.

3.2 Prediction Performance for Unbound Structures

In the previous subsection, 5-fold cross validation was performed using bound protein structures. However, prediction for unbound structures would be more useful for real applications. Thus, we examined the difference between the prediction for bound structures and the prediction for unbound structures. We searched PDB for unbound structures corresponding to bound structures of hetero complexes in the dataset described in "Materials and Methods". As a result, we obtained 21 unbound structures. Interaction sites of both bound and unbound structures were predicted using SVM trained by the bound structures that do not have the corresponding unbound structures. The two-stage SVM using PSSMs and ASAs of polar and non-polar atoms in a residue was used for prediction. MCC values for unbound structures were plotted against those for bound structures (As shown in **Fig. 4**, there are no serious differences between the bound and unbound predictions. The reason for this seems to be that the structural change between the bound and the unbound structures are small: Cα RMSD (root mean square deviation) is less than 2.0 angstrom in almost (19/21) bound-unbound pairs.

3.3 Propensities of Interaction Sites

We calculated the ratio of interaction site residues in particular amino acids in protein surfaces. The results are shown in **Fig. 5**. In this figure, amino acid types are arranged in ascending order of hydropathic index[33]. This figure clearly shows that the ratio of interaction sites in hydrophobic amino acids is higher than

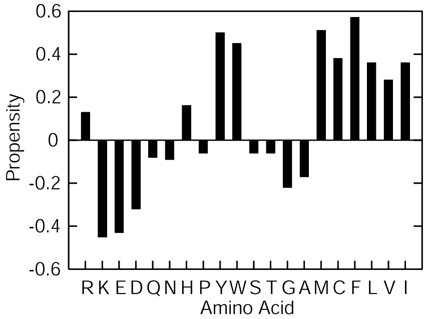

Fig. 5 Propensities of each amino acid type to be an interaction site. Propensities were calculated as base two logarithms. Positive value means that a residue appears in interaction site more frequently than on surface, and negative value means that a residue appears in interaction site less frequently than on surface.

in other amino acids. This is consistent with the work of Bordner, et al.[22] who showed that solvation energy calculated using atomic solvation parameters weighted by atomic ASA is high in the interface. Tryptophan and tyrosine are less hydrophobic but have a high propensity to be interaction sites. These residues are known for their aromatic side-chain interactions. Specifically, it is known that tryptophan often becomes an anchor residue, a hot spot that contributes greatly to free energy for binding[34], and the structural conservation of tryptophan on the protein surface indicates highly possible binding sites[35],[36].

To investigate how various properties discriminate between interaction sites and non interaction sites, we calculated the prediction performance of the hydropathy index (H), the hydropathy index weighted by relative ASA (HW), amino acid propensities (A), and amino acid propensities weighted by relative ASA (AW). These methods are statistical methods; they made predictions using the sum of the value of the central residue and 14 spatially neighboring residues. The threshold value of these statistical methods was adjusted by maximizing MCC. The results of these methods are summarized in **Table 3**. The table shows that 1) it is important to consider the exposure of residue, 2) amino acid propensities were a better basis on which to discriminate between interaction and non-interaction sites than the hydropathy index. First, the larger the exposure of residue is, the larger the energy stabilization by interaction is. Thus, exposure of residue is important for discrimination. Second, the hydropathy index cannot discriminate well because a low hydropathy index is assigned to weak hydrophilic residues such as tryptophan and tyrosine and strong hydrophilic residues such as arginine, but these residues have a strong propensity to be interaction sites (Fig. 5). As a comparison, we also present the result of two SVM results: One is the two-stage SVM with amino acid frequencies (propensities) and relative residue-base ASAs, and the other is the two-stage SVM with the PSSMs and relative residue-base ASAs. The result that the two SVM methods performed better than the statistical methods show that the machine learning technique (SVM) is effective to learn and predict the complicated pattern of the interaction sites. The reason why the PSSMs contribute to the performance improvement is that it takes into consideration evolutionary conservation and combination of amino acid types in patches rather than simple amino acid propensities.

Table 3 Discrimination ability of hydropathy index, amino acid propensities, and SVM method.

Method	Precision	Recall	MCC
H	41.3	56.4	0.199
HW	42.9	60.7	0.235
A	42.4	59.5	0.224
SVM-freq	49.2	72.6	0.368
SVM-PSSM	50.5	73.6	0.389

H: hydropathy index. HW: hydropathy index weighted by relative ASA. A: amino acid propensities. AW: amino acid propensities weighted by relative ASA. SVM-freq: two-stage SVM with frequencies and ASAs of polar and non-polar atoms in a residue. SVM-PSSM: two-stage SVM with PSSMs and ASAs of polar and non-polar atoms in a residue.

3.4 Example of Prediction

An example of prediction of interaction sites is shown in **Fig. 6** and **Table 4**. The three images are answer and prediction results for the catalytic domain of Ras GTPase-activating protein (PDB ID: 1wq1, chain ID: g). Top is for answer, middle is for two-stage SVM with PSSMs and ASAs of polar and non-polar atoms in a residue, and bottom is for single-stage SVM with PSSMs and ASAs of polar and non-polar atoms in a residue. The number of false positives predicted by the two-stage method was lower than that predicted by the single-stage method. The two-stage method for filtering worked well in this example. The hydrophobic region of the interaction site of this protein,

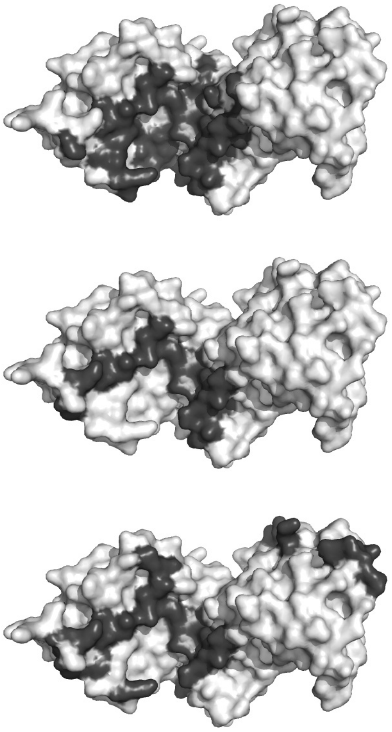

Fig. 6 Prediction results for 1wq1g. Top, middle, and bottom represent answer, result for two-stage SVM method, and result for single-stage SVM method, respectively. Black indicates interaction residue or predicted interaction residue, and white indicates non-interaction residue or predicted non-interaction residue.

Table 4 Prediction results for RasGAP (PDB ID: 1wq1g).

Method	Precision	Recall	MCC
Single-stage SVM	55.7	46.9	0.443
Two-stage SVM	88.2	46.9	0.602

which is rich in residues with a strong propensity to interact (Arg789, Leu902, Arg903, and Leu910), was correctly predicted to be an interaction site. Meanwhile the hydrophilic region of interaction site of this protein, which is rich in residues with a weak propensity to interact (Lys935, Gln938, Asn942, Lys949 and Glu950), was incorrectly predicted to be a non-interaction site. However, this hydrophilic region plays an important role in the conformational changes required for GTP hydrolysis[37]. Since the SVM learns an optimal hyperplane from the training dataset, which has propensities for interaction sites that are rich in hydrophobic and aromatic residues and are not rich in hydrophilic residues, it is difficult to correctly predict hydrophilic interaction sites using this SVM predictor.

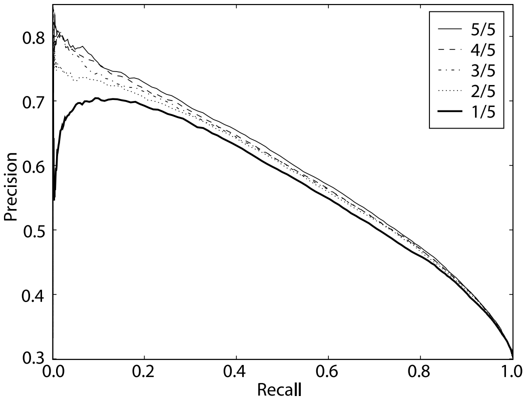

Fig. 7 Recall-precision curves of five predictors trained by subset of dataset.

3.5 Effect of Size of Training Dataset

To estimate how much performance improved as the size of the training dataset increased, we investigated the performances of five predictors trained by a subset of the dataset. Predictors were trained by 1/5, 2/5, 3/5, 4/5, and 5/5 of the dataset described in "Materials and Methods". These predictors used the two-stage SVM with PSSMs and ASAs of polar and non-polar atoms. The results are shown in **Fig. 7**. The performances of these five predictors increase as the size of the dataset increases. However, the improvements in performance become progressively smaller. This suggests that SVM learning to predict interaction sites from PSSMs and ASAs is reaching a limit. Since the number of non-homologous complex structures is unlikely to increase rapidly, the improvement of performance based on growth of PDB is expected to be small.

4. Conclusion

We investigated the propensities of protein interaction site residues and developed a novel prediction method using results of the investigation. In the dataset we used, hydrophobic and aromatic residues are highly likely interaction site residues. We developed a method using sequence information alone and a method using both sequence and structural information. Thus, we can predict protein-protein interaction sites from sequence alone, and, if the structure of target protein is available, it is possible to predict that more accurately. PSSMs, ASAs, and two-stage SVM all contributed to improvement in performance. The contribution of PSSMs to improvement was the greatest.

Performance comparison with other methods described in this paper is restricted. The existing systems have evaluated performance for different training and prediction sets and it is practically difficult to fairly compare performance among them.

We are planning to further modify our method and to apply it to docking algorithms. The prediction of interaction residues may be useful for restriction of search spaces or for filtering out incorrect predictions. Many methods of predicting interaction sites have been developed. Most of them pay attention only to information about target proteins. When a partner protein is known, using the information about a partner such as amino acid composition, presence or absence of a hydrophobic patch or clusters of charged residues may possibly improve prediction performance. We will also use this method for docking algorithms developed in our laboratory[38].

Acknowledgments This work was supported in part by Grants-in-Aid for Scientific Research on Priority Areas (No.16014204) from the Ministry of Education, Culture, Sports, Science and Technology of Japan to K.S.

References

1) Fields, S. and Song, O.: A novel genetic system to detect protein-protein interactions, *Nature*, Vol.340, No.6230, pp.245–246 (1989).

2) Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, Vol.403, No.6770, pp.623–627 (2000).

3) Halperin, I., Ma, B., Wolfson, H. and Nussinov, R.: Principles of docking: An overview of search algorithms and a guide to scoring functions, *Proteins*, Vol.47, No.4, pp.409–443 (2002).

4) Wodak, S.J. and Mendez, R.: Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications, *Curr Opin Struct Biol*, Vol.14, No.2, pp.242–249 (2004).

5) Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J.E., Vajda, S., Vakser, I. and Wodak, S.J.: CAPRI: A Critical Assessment of PRedicted Interactions, *Proteins*, Vol.52, No.1, pp.2–9 (2003).

6) de Vries, S.J., van Dijk, A.D.J. and Bonvin, A.M.J.J.: WHISCY: what information does surface conservation yield? Application to data-driven docking, *Proteins*, Vol.63, No.3, pp.479–489 (2006).

7) Jones, S. and Thornton, J.M.: Analysis of protein-protein interaction sites using surface patches, *J Mol Biol*, Vol.272, No.1, pp.121–132 (1997).

8) Ofran, Y. and Rost, B.: Analysing six types of protein-protein interfaces, *J Mol Biol*, Vol.325, No.2, pp.377–387 (2003).

9) Valdar, W.S. and Thornton, J.M.: Protein-protein interfaces: Analysis of amino acid conservation in homodimers, *Proteins*, Vol.42, No.1, pp.108–124 (2001).

10) Jones, S. and Thornton, J.M.: Prediction of protein-protein interaction sites using patch analysis, *J Mol Biol*, Vol.272, No.1, pp.133–143 (1997).

11) Neuvirth, H., Raz, R. and Schreiber, G.: ProMate: A structure based prediction program to identify the location of protein-protein binding sites, *J Mol Biol*, Vol.338, No.1, pp.181–199 (2004).

12) Fernandez-Recio, J., Totrov, M., Skorodumov, C. and Abagyan, R.: Optimal docking area: A new method for predicting protein-protein interaction sites, *Proteins*, Vol.58, No.1, pp.134–143 (2005).

13) Lichtarge, O., Bourne, H.R. and Cohen, F.E.: An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol*, Vol.257, No.2, pp.342–358 (1996).

14) Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L. and Lichtarge, O.: An accurate, sensitive, and scalable method to identify functional sites in protein structures, *J Mol Biol*, Vol.326, No.1, pp.255–261 (2003).

15) Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J.: Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking, *J Mol Biol*, Vol.311, No.2, pp.395–408 (2001).

16) Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N.: Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bioinformatics*, Vol.18, Suppl. 1, pp.S71–S77 (2002).

17) Armon, A., Graur, D. and Ben-Tal, N.: ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface

mapping of phylogenetic information, *J Mol Biol*, Vol.307, No.1, pp.447–463 (2001).

18) Ofran, Y. and Rost, B.: Predicted protein-protein interaction sites from local sequence information, *FEBS Lett*, Vol.544, No.1-3, pp.236–239 (2003).

19) Koike, A. and Takagi, T.: Prediction of protein-protein interaction sites using support vector machines, *Protein Eng Des Sel*, Vol.17, No.2, pp.165–173 (2004).

20) Fariselli, P., Pazos, F., Valencia, A. and Casadio, R.: Prediction of protein–protein interaction sites in heterocomplexes with neural networks, *Eur J Biochem*, Vol.269, No.5, pp.1356–1361 (2002).

21) Zhou, H.X. and Shan, Y.: Prediction of protein interaction sites from sequence profile and residue neighbor list, *Proteins*, Vol.44, No.3, pp.336–343 (2001).

22) Bordner, A.J. and Abagyan, R.: Statistical analysis and prediction of protein-protein interfaces, *Proteins*, Vol.60, No.3, pp.353–366 (2005).

23) Bradford, J.R. and Westhead, D.R.: Improved prediction of protein-protein binding sites using a support vector machines approach, *Bioinformatics*, Vol.21, No.8, pp.1487–1494 (2005).

24) Hua, S. and Sun, Z.: A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *J Mol Biol*, Vol.308, No.2, pp.397–407 (2001).

25) Nguyen, M.N. and Rajapakse, J.C.: Prediction of protein relative solvent accessibility with a two-stage SVM approach, *Proteins*, Vol.59, No.1, pp.30–37 (2005).

26) Ding, C.H. and Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, Vol.17, No.4, pp.349–358 (2001).

27) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E.: The Protein Data Bank, *Nucleic Acids Res*, Vol.28, No.1, pp.235–242 (2000).

28) Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res*, Vol.25, No.17, pp.3389–3402 (1997).

29) Kabsch, W. and Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, Vol.22, No.12, pp.2577–2637 (1983).

30) Hubbard, S. and Thornton, J.: NACCESS, Technical report, Department of Biochemistry and Molecular Biology, University College London (1993).

31) Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol.20, No.3, pp.273–297 (1995).

32) Chang, C.-C. and Lin, C.-J.: *LIBSVM: A library for support vector machines* (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

33) Kyte, J. and Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, Vol.157, No.1, pp.105–132 (1982).

34) Bogan, A.A. and Thorn, K.S.: Anatomy of hot spots in protein interfaces, *J Mol Biol*, Vol.280, No.1, pp.1–9 (1998).

35) Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R.: Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proc Natl Acad Sci U S A*, Vol.100, No.10, pp.5772–5777 (2003).

36) Hu, Z., Ma, B., Wolfson, H. and Nussinov, R.: Conservation of polar residues as hot spots at protein interfaces, *Proteins*, Vol.39, No.4, pp.331–342 (2000).

37) Scheffzek, K., Ahmadian, M., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F. and Wittinghofer, A.: The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants., *Science*, Vol.277, No.5324, pp.333–338 (1997).

38) Sumikoshi, K., Terada, T., Nakamura, S. and Shimizu, K.: A fast protein-protein docking algorithm using series expansion in terms of spherical basis functions, *Genome Inform*, Vol.16, No.2, pp.161–173 (2005).

(Received September 12, 2007)
(Accepted December 17, 2007)

(Communicated by *Tatsuya Akutsu*)

Masanori Kakuta was born in 1982. He received his Master of Agrigulture degree from the University of Tokyo in 2007. Since 2007, he has been a graduate student at the University of Tokyo. His current research interest are protein-protein interactions and interaction sites. He is a member of BSJ and PSSJ.

Shugo Nakamura was born in 1968. He received his Ph.D. degree from the University of Tokyo in 2001. In the University of Tokyo, he had been an assistant professor since 1995 and has been an associate professor since 2002. His current research interests is structural and functional analysis of proteins and nucleic acids using computer, especially prediction of their structures from their sequences. He is a member of BSJ, PSSJ, and JSBi.

Kentaro Shimizu was born in 1957. He received his M.Sc. and D.Sc. degrees from the University of Tokyo in 1982 and 1985 respectively. He had been an assistant professor in the University of Tokyo and had engaged in research on operating systems. Since 1991 he had been in University of Electro-Communications as an associate professor and has been in the University of Tokyo as a professor since 1998. His current research interests are protein structure prediction, protein interaction prediction, docking simulation, and folding simulation. He is a member of IPSJ, JSBi, IEEE-CS, ACM, ACS, IEICE, BSJ, JS-BBA, and PSSJ.