Published online 4 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 9 ***3025–3030***
doi:10.1093/nar/gkn159

Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences

Yanzhi Guo[1,2], Lezheng Yu[1,2], Zhining Wen[1,2] and Menglong Li[1,2,*]

[1]College of Chemistry, Sichuan University, Chengdu 610064 and [2]State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, P.R. China

Received January 10, 2008; Revised March 3, 2008; Accepted March 20, 2008

ABSTRACT

Compared to the available protein sequences of different organisms, the number of revealed protein–protein interactions (PPIs) is still very limited. So many computational methods have been developed to facilitate the identification of novel PPIs. However, the methods only using the information of protein sequences are more universal than those that depend on some additional information or predictions about the proteins. In this article, a sequence-based method is proposed by combining a new feature representation using auto covariance (AC) and support vector machine (SVM). AC accounts for the interactions between residues a certain distance apart in the sequence, so this method adequately takes the neighbouring effect into account. When performed on the PPI data of yeast *Saccharomyces cerevisiae*, the method achieved a very promising prediction result. An independent data set of 11 474 yeast PPIs was used to evaluate this prediction model and the prediction accuracy is 88.09%. The performance of this method is superior to those of the existing sequence-based methods, so it can be a useful supplementary tool for future proteomics studies. The prediction software and all data sets used in this article are freely available at http://www.scucic.cn/Predict_PPI/index.htm.

INTRODUCTION

Identification of protein–protein interactions (PPIs) is crucial for elucidating protein functions and further understanding various biological processes in a cell. It has been the focus of the post-proteomic researches. In recent years, various experimental techniques have been developed for the large-scale PPI analysis, including yeast two-hybrid systems (1,2), mass spectrometry (3,4), protein chip (5) and so on. Because experimental methods are time-consuming and expensive, current PPI pairs obtained from experiments only cover a small fraction of the complete PPI networks (6). Hence, it is of great practical significance to develop the reliable computational methods to facilitate the identification of PPIs.

So far, a number of computational methods have been proposed for the prediction of PPIs. Some methods are based on the genomic information, such as phylogenetic profiles (7), gene neighbourhood (8) and gene fusion events (9,10). Methods using the structural information of proteins (11–13) and the sequence conservation between interacting proteins (14,15) have been reported. Previously predicted (known protein) domains that are responsible for the interactions between proteins have also been considered too (16–20). However, all these methods cannot be implemented if such pre-knowledge about the proteins is not available. Several sequence-based methods (21–25) have shown that the information of amino acid sequences alone may be sufficient to identify novel PPIs, but the highest accuracy of these methods is only ~80%, such as the methods by Martin *et al*. (22), and Chou and Cai (25). So Shen *et al*. (26) have developed an alternative method that yields a high prediction accuracy of 83.9%, when applied to predicting human PPIs. This method considers the local environments of residues through a conjoint triad method, but it only accounts for the properties of one amino acid and its proximate two amino acids. However, the interactions usually occur in the discontinuous amino acids segments in the sequence, and the information of these interactions may be able to further improve the prediction ability of the existing sequence-based methods.

In this article, a new method based on support vector machine (SVM) and auto covariance (AC) was proposed. AC accounts for the interactions between amino acids within a certain number of amino acids apart in the sequence, so this method takes neighbouring effect into

*To whom correspondence should be addressed. Tel: +86 28 89005151; Fax: +86 28 85412356; Email: liml@scu.edu.cn

account and makes it possible to discover patterns that run through entire sequences. The amino acid residues were translated into numerical values representing physicochemical properties, and then these numerical sequences were analysed by AC based on the calculation of covariance. Finally, the SVM model was constructed using the vectors of AC variables as input. The optimization experiment demonstrated that the interactions of one amino acid and its 30 vicinal amino acids would contribute to characterizing the PPI information. The method was tested by the PPI data of yeast *Saccharomyces cerevisiae* and yielded a prediction accuracy of 87.36%. At last, this model was further evaluated by an independent data set of other yeast PPIs with the prediction accuracy of 88.09%.

MATERIALS AND METHODS

Data collection and data set construction

The PPI data was collected from *Saccharomyces cerevisiae* core subset of database of interacting proteins (DIP) (27), version DIP_20070219. The reliability of this core subset has been tested by two methods, expression profile reliability (EPR) and paralogous verification method (PVM) (28). At the time of doing the experiments, the core subset contained 5966 interaction pairs. The protein pairs that contained a protein with <50 amino acids were removed and the remaining 5943 protein pairs comprised the final positive data set. All proteins in the data set were aligned using the multiple sequence alignment tool, cd-hit program (29). The aligned result shows that among the 5943 protein pairs, the overwhelming majority of them (5594 PPIs) have <40% pairwise sequence identity to one another. Although there are only 349 pairs with ≥40% identity in the training data set, the classifier will possibly be biased to these homologous sequence pairs.

Since the non-interacting pairs were not readily available, three strategies for constructing negative data set were used in order to compare the effects of different training data sets on the performance of the method. The first strategy has been described by Shen and colleagues (26) in detail. The non-interacting pairs were generated by randomly pairing proteins that appeared in the positive data set. Here the negative data set based on this method is called Prcp. The second is based on such an assumption that proteins occupying different subcellular localizations do not interact. The subcellular localization information of the proteins in the positive data set was extracted from Swiss-Prot (http://www.expasy.org/sprot/). The proteins without subcellular localization information and those denoted as 'putative', 'hypothetical' were excluded. The remaining proteins were grouped into eight subsets based on the eight main types of localization—cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, golgi apparatus, peroxisome, vacuole and cytoplasm&nucleus. Each subset contained 10 proteins at least. The non-interacting pairs were generated by pairing proteins from one subset with proteins from the other subset. It must be pointed out that proteins from cytoplasm subset and nucleus subset cannot be paired with those from cytoplasm&nucleus subset. Here the negative data set based on subcellular localization information is called Psub. The two strategies must meet three requirements: (i) the non-interacting pairs cannot appear in the whole DIP yeast interacting pairs, (ii) the number of negative pairs is equal to that of positive pairs and (iii) the contribution of proteins in negative set should be as harmonious as possible (24,26).

As a comparison, the third strategy was used for creating non-interacting pairs composed of artificial protein sequences. It has been demonstrated that if a sequence of one interacting pair is shuffled, then the two proteins can be deemed not to interact with each other (30). Thus, the negative data set was prepared by shuffling the sequences of right-side interacting pairs with k-let (k = 1,2,3) counts using the Shuffle program (31).

Feature extraction and AC

Protein–protein interaction can be defined as four interaction modes: electrostatic interaction, hydrophobic interaction, steric interaction and hydrogen bond. Here seven physicochemical properties of amino acids were selected to reflect these interaction modes whenever possible and they are hydrophobicity (32), hydrophicility (33), volumes of side chains of amino acids (34), polarity (35), polarizability (36), solvent-accessible surface area (SASA) (37) and net charge index (NCI) of side chains of amino acids (38), respectively. The original values of the seven physicochemical properties for each amino acid are listed in Supplementary Table S1. They were first normalized to zero mean and unit standard deviation (SD) according to Equation (1):

$$P'_{ij} = \frac{P_{i,j} - P_j}{S_j} \quad \mathbf{1}$$

where $P_{i,j}$ is the j-th descriptor value for i-th amino acid, P_j the mean of j-th descriptor over the 20 amino acids and S_j the corresponding SD. Then each protein sequence was translated into seven vectors with each amino acid represented by the normalized values of seven descriptors.

Artificial intelligence-based techniques such as SVM and the neural network require a fixed number of inputs for training. However, there are often unequal-length vectors because of protein sequences with different lengths. So auto cross covariance (ACC) was used to transform these numerical vectors into uniform matrices. As a statistical tool for analyzing sequences of vectors developed by Wold *et al*. (39), ACC has been adopted by more and more leading investigators for protein classification (40–42). ACC results in two kinds of variables, AC between the same descriptor, and cross covariance (CC) between two different descriptors. In this study, only AC variables were used in order to avoid generating too large number of variants, compared to the limited number of PPI pairs. Given a protein sequence, AC variables describe the average interactions between residues, a certain *lag* apart throughout the whole sequence. Here, *lag* is the distance between one residue and its neighbour, a certain number of residues away. The AC variables are calculated according to Equation (2), where j represents

one descriptor, i the position in the sequence X, n the length of the sequence X and lag the value of the lag.

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \times \left(X_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \quad \mathbf{2}$$

In this way, the number of AC variables, D can be calculated as $D = lg \times P$, where P is the number of descriptors and lg is the maximum lag ($lag = 1, 2, \ldots, lg$). After each protein sequence was represented as a vector of AC variables, a protein pair was characterized by concatenating the vectors of two proteins in this protein pair.

Model construction

The classification model for predicting PPIs was based on SVM. Vapnik (43) has given a full description about how to use SVM to do classification. The software libsvm 2.84 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was employed in this work. A radial basis function (RBF) was chosen as the kernel function. Two parameters, the regularization parameter C and the kernel width parameter γ were optimized using a grid search approach. In statistical prediction, sub-sampling test and jackknife test are often used as two cross-validation methods (44). Jackknife test is deemed more objective and has been widely adopted by many investigators (41,45–56) to test the power of various predictors, but it will take much long time to perform the jackknife test. Although it has been demonstrated that the sub-sampling test cannot avoid arbitrariness according to a recent comprehensive review (45) and a penetrating analysis in (57), it is still a good validation method for the large data set. Considering the numerous samples used in this work, 5-fold cross-validation was used to investigate the training set.

The final data set consisted of 11 886 protein pairs, half from the positive data set and half from the negative data set. Here three-fifths of the protein pairs respectively from the positive and negative data set were randomly chosen as the training set (7130 protein pairs) and the remaining two-fifths (4576 protein pairs) were used as the test set. An SVM model was built using the training set and 5-fold cross-validation, and the performance of this model was evaluated by the test set. In order to test the robustness of the method, this process of random selection of training set and test set, model-building and model-evaluating was repeated five times. Thus, five training sets and five test sets were prepared, so five models were generated. Three parameters, sensitivity, precision and accuracy were used to measure the performance of this method. They are defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \mathbf{3}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \mathbf{4}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \mathbf{5}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

RESULTS AND DISCUSSION

Comparing the prediction performances of different negative data sets

SVM models were constructed using the five negative data sets derived from the three strategies. In this step, lg was initialized to be 25 amino acids. Table 1 gives the average prediction results of SVM models using different negative data sets. Using 1-let, 2-let and 3-let shuffled protein sequences as negative data set, the average prediction accuracy is 79.25, 77.30 and 70.25%, respectively. The trend that the prediction accuracy decreases with the increase of k is resulted from the fact that the shuffling procedure provides more native-like artificial proteins by conserving higher-order biases. The model based on the negative data set Prcp yields very low prediction accuracy of 58.42%, while the model built with the same strategy by Shen *et al.* (26) achieves a good performance with an accuracy of 83.90%. It is probably due to the different feature representation methods and data sources. Compared to other four models, the model based on the negative data set Psub gives the best performance. The average prediction accuracy, sensitivity and precision are 86.23, 85.22 and 87.83%, respectively, which indicate that this method is successful in predicting PPIs using the non-interacting pairs of non co-localized proteins as the negative data set. However, it is necessary to point out that selecting non-interacting pairs of non co-localized protein will lead to over-optimistic estimates of classifier accuracy, as denoted by Ben-hur and Noble (58).

Selecting optimal *lg*

The use of AC with large *lag*s will result in more variables that account for interactions of amino acids with large

Table 1. The comparative results of the prediction performance of the method based on different negative data sets, respectively, using AC with lg of 25 amino acids

Negative data set	Psub	Prcp	1-let	2-let	3-let
Sensitivity (%)	85.22	41.76	79.29	69.81	60.74
Precision (%)	87.83	62.64	82.67	85.14	80.15
Accuracy (%)	86.23 ± 1.95	58.42 ± 1.68	79.25 ± 7.80	77.30 ± 12.38	70.25 ± 10.40

Psub is the negative data set of non-interacting pairs of non-co-localized proteins; Prcp is the negative data set derived from the method by Shen *et al.* (26). The three negative data sets, 1-let, 2-let and 3-let are obtained by shuffling the protein sequences with k-let counts, $k = 1, 2, 3$.

distances apart in the sequence. The maximal possible *lg* is the length of the shortest sequence (50 amino acids) in the data set. In this study, several *lg*s were optimized in order to achieve the best characterization of the protein sequences. Using Psub as the negative data set, nine models were constructed with nine different *lg*s, respectively (*lg* = 5, 10, 15, 20, 25, 30, 35, 40, 45). The prediction results for the nine models are shown in Figure 1. As seen from the curve, the prediction accuracy increases when *lg* increases from 5 to 30, but it slightly fluctuates when *lg* increases from 30 up to 45. There is a peak point with an average accuracy of 87.36% and the *lg* of 30 amino acids. It is concluded that AC with *lg* less than 30 amino acids would lose some useful features of the protein sequences and larger *lg*s could introduce noise instead of improving the prediction power of the model. So the optimal *lg* is 30 amino acids.

Comparing the performance of AC with that of ACC

After represented by the seven descriptors, a protein pair was converted into a 420-dimensional ($2 \times 30 \times 7$) vector by AC with *lg* of 30 amino acids. However, when ACC is used, a protein sequence will be a vector of 2940 dimension ($2 \times 30 \times 7 \times 7$). To reduce the calculating time, only AC variables were used as the input of SVM. Here, we also used ACC to transform the protein sequences and compared the performance of the model based on ACC with that of the model based on AC. From Table 2, we can see that the model based on ACC transform gives good results with the average sensitivity, precision and accuracy of 89.93, 88.87 and $89.33 \pm 2.67\%$, respectively. However, when the dimension of vector space is dramatically reduced from 2940 to 420 using AC transform, the performance of the model based on AC is very close to that of the model based on ACC. It proves that CC variables only have a little contribution to the performance of the model and AC variables are the principal components of ACC variables.

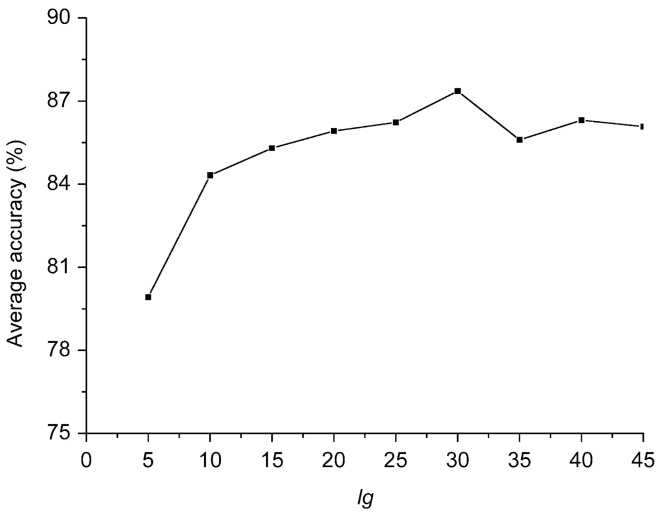

Figure 1. The average prediction accuracy of the method with AC of different *lg*s respectively.

So in this work, the optimal model was based on the negative data set Psub and AC transform with *lg* of 30 amino acids. The prediction results for five test sets are listed in Table 2. For all five models, the prediction accuracies are all >86% with a relatively low SD of 1.38%. On average, the sensitivity, precision and prediction accuracy of this model are 87.30, 87.82 and 87.36%, respectively. These results are obtained based on the original data set that contains homologous protein pairs. However, for the statistical predictions, it is absolutely necessary to avoid redundancy and homology bias in the training data set (57). In order to determine the homology effects, the non-redundant data set was constructed by removing the protein pairs with ≥40% pairwise sequence identity from the whole original data set. The performance of the five models based on this non-redundant data set is shown in Supplementary Table S2. The average prediction accuracy of the non-redundant data set is 86.55%.

Two SVM parameters, C and γ were optimized as 32 and 0.03125. So using the whole data set, the final prediction model was built with the optimal parameters.

Performance on the independent data set

In order to evaluate the practical prediction ability of the final prediction model, a large independent data set was constructed. In DIP, the yeast data set contained 17 491 interaction pairs, out of which that which contained

Table 2. The prediction results of the test sets based on the negative data set Psub and *lg* of 30 amino acids

	Test set	TP	FN	TN	FP	Sensitivity (%)	Precision (%)	Accuracy (%)
ACC	1	2096	282	2226	152	88.14	93.24	90.87
	2	2282	96	1741	637	95.96	78.18	84.59
	3	2023	355	2291	87	85.07	95.88	90.71
	4	2181	197	2099	279	91.72	88.66	89.99
	5	2052	267	2194	184	88.77	91.98	90.52
	Average	2138	240	2110	268	89.93	88.87	89.33 ± 2.67
AC	1	2161	217	1944	434	90.87	83.28	86.31
	2	2215	163	1890	488	93.15	81.95	86.31
	3	2062	316	2153	225	86.71	90.16	88.63
	4	1890	488	2221	157	79.48	92.33	86.44
	5	2052	326	2185	193	86.29	91.40	89.10
	Average	2076	312	2079	299	87.30	87.82	87.36 ± 1.38

TP, true positive; FP, false positive; TN, true negative; FN, false negative; Psub is the negative data set of non-interacting pairs of non-co-localized proteins.

a protein with <50 amino acids and those appearing in the training data set were all excluded. Among the remaining 11 474 protein pairs, 10 108 PPIs are correctly predicted by the prediction model and the success rate is 88.09%. In this article, the negative training set was generated by selecting non-interacting pairs of non-co-localized proteins. However, Ben-Hur and Noble (58) have denoted that restricting negative examples to non-co-localized protein pairs leads to a biased estimate of the accuracy of a PPI predictor. So it is necessary to generate a test data set of the non-interacting pairs with the same localization to test the effects of this bias. The yeast proteins used in the positive training set were assigned with the seven main types of localization. The non-interacting protein pairs with the same localization were generated and none of them has occurred in the whole DIP yeast interacting pairs. The performance of this method in predicting such negative samples is summarized in Supplementary Table S3. For cytoplasm and nucleus subsets, only 8000 non-interactions were randomly selected from the large-scale data set, respectively. The result shows that the prediction model is able to correctly predict the non-interacting pairs of all subsets with >80% accuracy, except the cytoplasm subset with 77% accuracy and endoplasmic reticulum subset with 69% accuracy. For all 27 204 non-interactions, the total prediction accuracy is 81.46%. In addition, using the model based on the non-redundant data set, the prediction accuracy for 11 474 yeast PPIs is 93.25% and the result of the non-interacting pairs is shown in Supplementary Table S4. All these results demonstrate that this method is also able to predict non-interacting pairs with the same localization.

CONCLUSION

In this article, we developed a new method for predicting PPIs only using the primary sequences of proteins. The prediction model was constructed based on SVM and AC. Shen *et al*. (26) have denoted that usually the methods with no local environments of amino acids are not reliable and robust, so they proposed a conjoint triad method to consider the properties of each amino acid and its two proximate amino acids. However, in most cases, the long-range interactions are also important for representing the PPI information. In this article, AC was used to involve the information of interactions between amino acids a longer distance apart in the sequence. A protein sequence was characterized by a series of ACs that covered the information of interactions between one amino acid and its 30 vicinal amino acids in the sequence. So this method adequately takes the neighbouring effect into account. As expected, this method improved the prediction accuracy compared with the current methods. Moreover, three different negative data sets were compared and the model trained using non-interacting pairs of non co-localized proteins yielded the best performance with a high accuracy of 87.36%, when applied to predicting the PPIs of *S. cerevisiae*. Meanwhile, the final prediction model was tested using the independent data set of the yeast PPIs with a good performance. Overall, such a robust method will be a useful tool to elucidate the biological function of newly discovered proteins and to expedite the study of protein networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors gratefully thank Eivind Coward for sharing the Shufflet sequence-randomizing code. The work was funded by the National Natural Science Foundation of China (No. 20775052). Funding to pay the Open Access publication charges for this article was provided by the National Natural Science Foundation of China (No. 20775052).

Conflict of interest statement. None declared.

REFERENCES

1. Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.
2. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
3. Gavin,A.C., Boche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J., Michon,A. and Cruciat,C. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
4. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
5. Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, **193**, 2101–2105.
6. Han,J.D., Dupuy,D., Bertin,N., Cusick,M.E. and Vidal,M. (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.*, **23**, 839–844.
7. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
8. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.*, **1**, 93–108.
9. Marcotte,E.M. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
10. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
11. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
12. Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.
13. Ogmen,U., Keskin,O., Aytuna,A.S., Nussinov,R. and Gursoy,A. (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Res.*, **33**, W331–W336.
14. Huang,T.W., Tien,A.C., Huang,W.S., Lee,Y.C., Peng,C.L., Tseng,H.H., Kao,C.Y. and Huang,C.Y. (2004) POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–3276.
15. Espadaler,J., Romero-Isart,O., Jackson,R.M. and Oliva,B. (2005) Prediction of protein–protein interactions using distant conservation

of sequence patterns and structure relationships. *Bioinformatics*, **21**, 3360–3368.
16. Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.*, **311**, 681–692.
17. Kim,W.K., Park,J. and Suh,J.K. (2002) Large scale statistical prediction of protein–protein interaction by potentially interacting domain (PID) pair. *Genome Inform.*, **13**, 42–50.
18. Han,D.S., Kim,H.S., Jang,W.H., Lee,S.D. and Suh,J.K. (2004) PreSPI: a domain combination based prediction system for protein–protein interaction. *Nucleic Acids Res.*, **32**, 6312–6320.
19. Morrison,J.L., Breitling,R., Higham,D.J. and Gilbert,D.R. (2006) A lock-and-key model for protein–protein interaction. *Bioinformatics*, **22**, 2212–2019.
20. Singhal,M. and Resat,H. (2007) A domain-based approach to predict protein–protein interactions. *BMC Bioinformatics*, **8**, 199.
21. Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
22. Martin,S., Roe,D. and Faulon,J.L. (2005) Predicting protein–protein interactions using signature products. *Bioinformatics*, **21**, 218–226.
23. Lo,S.L., Cai,C.Z., Chen,Y.Z. and Chung,M.C.M. (2005) Effect of training datasets on support vector machine prediction of protein–protein interactions. *Proteomics*, **5**, 876–884.
24. Pitre,S., Dehne,F., Chan,A., Cheetham,J., Duong,A., Emili,A., Gebbia,M., Greenblatt,J., Jessulat,M., Krogan,N. *et al.* (2006) PIPE: a protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.
25. Chou,K.C. and Cai,Y.D. (2006) Predicting protein–protein interactions from sequences in a hybridization space. *J. Proteome Res.*, **5**, 316–322.
26. Shen,JW., Zhang,J., Luo,X.M., Zhu,W.L., Yu,K.Q., Chen,K.X., Li,Y.X. and Jiang,H.L. (2007) Predicting protein–protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.
27. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP: the database of interacting proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.
28. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.
29. Li,W.Z., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
30. Kandel,Y., Matias,R., Unger,R. and Winkler,P.M. (1996) Shuffling biological sequences. *Discrete Appl. Math.*, **71**, 171–185.
31. Coward,E. (1999) Shufflet: shuffling sequences while conserving the *k*-let counts. *Bioinformatics*, **15**, 1058–1059.
32. Tanford,C. (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.*, **84**, 4240–4274.
33. Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
34. Krigbaum,W.R. and Komoriya,A. (1979) Local interactions as a structure determinant for protein molecules: II. *Biochim. Biophys. Acta*, **576**, 204–228.
35. Grantham,R. (1974) Amino acid difference formular to help explain protein evolution. *Science*, **185**, 862–864.
36. Charton,M. and Charton,B.I. (1982) The structure dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.*, **99**, 629–644.
37. Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.
38. Zhou,P., Tian,F.F., Li,B., Wu,S.R. and Li,Z.L. (2006) Genetic algorithm-base virtual screening of combinative mode for peptide/protein. *Acta Chim. Sinica*, **64**, 691–697.
39. Wold,S., Jonsson,J., Sjöström,M., Sandberg,M. and Rännar,S. (1993) DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta*, **277**, 239–253.
40. Guo,Y.Z., Li,M.L., Lu,M.C., Wen,Z.N. and Huang,Z.T. (2006) Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. *Proteins*, **65**, 55–60.
41. Wen,Z.N., Li,M.L., Li,Y.Z., Guo,Y.Z. and Wang,K.L. (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. *Amino Acids*, **32**, 277–283.
42. Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, **8**, 4.
43. Vapnik,V. (1998) *Statistical learning theory* Wiley, New York.
44. Chou,K.C. and Zhang,C.T. (1995) Review: prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.
45. Chou,K.C. and Shen,H.B. (2007) Review: recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.
46. Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.
47. Huang,Y. and Li,Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, **20**, 21–28.
48. Du,P.F. and Li,Y.D. (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*, **7**, 518.
49. Mondal,S., Bhavna,R., Mohan-Babu,R. and Ramakumar,S. (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.*, **243**, 252–260.
50. Guo,J., Lin,Y.L. and Liu,X.J. (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. *Proteomics*, **6**, 5099–5105.
51. Kedarisetti,K.D., Kurgan,L. and Dick,S. (2006) Classifier ensembles for protein structural class prediction with varying homology. *Biochem. Biophys. Res. Commun.*, **348**, 981–988.
52. Guo,Y.Z., Li,M.L., Lu,M.C., Wen,Z.N., Wang,K.L., Li,G.B. and Wu,J. (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. *Amino Acids*, **30**, 397–402.
53. Zhang,T.L. and Ding,Y.S. (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. *Amino Acids*, **33**, 623–629.
54. Pugalenthi,G., Tang,K., Suganthan,P.N., Archunan,G. and Sowdhamini,R. (2007) A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. *BMC Bioinformatics*, **8**, 351.
55. Tan,F.Y., Feng,X.Y., Fang,Z., Li,M.L., Guo,Y.Z. and Jiang,L. (2007) Prediction of mitochondrial proteins based on genetic algorithm: partial least squares and support vector machine. *Amino Acids*, **33**, 669–675.
56. Diao,Y.B., Ma,D.C., Wen,Z.N., Yin,J.J., Xiang,J. and Li,M.L. (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, **34**, 111–117.
57. Chou,K.C. and Shen,H.B. (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **3**, 153–162.
58. Ben-Hur,A. and Noble,W.S. (2006) Choosing negative examples for the prediction of protein–protein interactions. *BMC Bioinformatics*, **7**, S2.