FEBS Letters 580 (2006) 380–384

Predicting protein interaction sites from residue spatial sequence profile and evolution rate

Bing Wang[a,b,1], Peng Chen[a,b,1], De-Shuang Huang[a,*], Jing-jing Li[a], Tat-Ming Lok[c], Michael R. Lyu[d]

[a] *Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China*
[b] *Department of Automation, University of Science and Technology of China, Hefei, Anhui 230026, China*
[c] *Information Engineering Department, The Chinese University of Hong Kong, Shatin, Hong Kong*
[d] *Computer Science and Engineering Department, The Chinese University of Hong Kong, Shatin, Hong Kong*

Received 17 October 2005; revised 29 November 2005; accepted 30 November 2005

Available online 19 December 2005

Edited by Robert B. Russell

Abstract **This paper proposes a novel method that can predict protein interaction sites in heterocomplexes using residue spatial sequence profile and evolution rate approaches. The former represents the information of multiple sequence alignments while the latter corresponds to a residue's evolutionary conservation score based on a phylogenetic tree. Three predictors using a support vector machines algorithm are constructed to predict whether a surface residue is a part of a protein–protein interface. The efficiency and the effectiveness of our proposed approach is verified by its better prediction performance compared with other models. The study is based on a non-redundant data set of heterodimers consisting of 69 protein chains.**

Keywords: Protein interaction sites; Support vector machines; Spatial sequence profile; Evolutionary rate; Multiple sequence alignments; Phylogenetic tree

1. Introduction

Protein–protein interactions play a critical role in live biological cells by controlling the functions that proteins perform, such as regulation of metabolic and signaling pathways, immunological recognition, DNA replication and gene translation, as well as protein synthesis [1]. Localization of such interactions to so-called ''functional sites'' or ''interaction sites'' will allow us to understand how the protein recognizes other molecules, to gain clues about its likely function at the level of the cell and the organism, and to identify important binding sites that may serve as useful targets for pharmaceutical design [2]. This is stimulating researchers to seek potential computational approaches for identifying the roles of function residues, especially those at protein–protein interaction sites.

Recently, a series of computational efforts to identify interaction sites or interfaces in proteins have been undertaken; these have addressed various aspects of protein structure and behavior, such as detecting the presence of ''proline brackets'' [3], solvent-accessible surface area buried upon association [4], free energy changes upon alanine-scanning mutations [5], in silico two hybrid systems [6], and sequence hydrophobicity distribution [7]. Jones and Thornton [8,9] successfully predicted protein interfaces by analyzing six parameters of surface patch. Also, in recent years, several studies have attempted to predict protein–protein interaction sites from sequence or structure conservation information [10–16].

These existing methods tackled the problem of protein–protein interaction from different angles, and the development of computational approaches to identify protein interaction sites is still at its embryonic stage.

In this paper, we present a novel, efficient method, which incorporates residue spatial sequence profile and evolution rate, to identify protein–protein interaction sites on the protein residue level. Amino acid sequence profile and evolution rate represent the information about evolutionary conservation base on multiple sequence alignments (MSAs) and the phylogenetic tree, respectively. Our purpose is to develop a general approach that can capture the general properties of interface residues, so we focus here on heterocomplexes for the reason that interacting surfaces in homocomplexes are characterized by hydrophobicity. To this end, a support vector machines (SVMs) predictor has been constructed for identifying protein interaction sites in protein chains. The results based on a non-redundant set of protein heterodimers demonstrated that this approach is effective and efficient; the model achieved a sensitivity of 66.3%, a specificity of 49.7%, an accuracy of 0.654 and a correlation coefficient of 0.297.

2. Materials and methods

2.1. Dataset preparation

To generate a predictor that can capture the general properties of residues located on a protein interface, we extracted a data source from a set of 113 pairs of interacting protein chains used in the study of Fariselli et al. [12]. The dataset eliminates homocomplexes and protease-inhibitor complexes, whose interacting surfaces are characterized by hydrophobicity and serine/histidine active site signatures, respectively. The dataset also excludes chains labeled as 'membrane peptides', 'small proteins' or 'coiled coils' in the SCOP classification [17]. After removal of redundant chains, we obtained a data set of 69 protein chains (sequence identity <30%); all the data are available upon request.

In this paper, a residue is considered to be a surface residue if its relative accessible surface area (ASA) is at least 16% of its nominal maximum area whose value as defined by Rost and Sander [18]. The ASA is computed for each residue in each protein chain using the DSSP program [19]. Here, we should emphasize that only the coordinates of the

*Corresponding author. Fax: +86 0551 5592420.
E-mail address: dshuang@iim.ac.cn (D.-S. Huang).

[1] B.W. and P.C. contributed equally to this work.

0014-5793/$32.00
doi:10.1016/j.febslet.2005.11.081

unbound chain were used in the calculation. If other chains present in the complex were included, their influence would cause the ASA to be incorrectly calculated. A residue is classified as an interface residue if the spatial distance between its α-carbon (CA) atom and random CA atoms in the other chains in the complex is less than 1.2 nm [12]. According to the above definitions, we obtain 10329 surface residues, 34.8% of which are interface residues.

2.2. Predictor construction

In our experiment, predictors are generated using the SVM algorithm to judge whether a residue is located on an interface or not. SVMs frequently demonstrate high prediction accuracy whilst avoiding over-fitting. They can also handle large feature spaces and condense the information given by the training dataset using support vectors [20]. Here, we consider only surface residues in the predictor training, the target value of which is 1 (positive sample) if the target residue is classified into the interface residue set and −1 (negative sample) otherwise. The SVM algorithm implemented here can be downloaded freely (http://www.cs.waikato.ac.nz/~ml/weka).

We constructed three SVM predictors using residue sequence profile, evolution rate, or a combination of these two attributes. For the predictor using residue sequence profile, the input vectors are obtained from the HSSP database [21], where each amino acid is represented by elements whose values are based on multiple alignments of protein sequences and their potential structural homologs. For the evolutionary rate [22–24] based predictor, each input vector is assigned a conservation score to amino acid position. Following the method used by Fariselli et al. [12], the input vector of these predictors is fed with a window of 11 residues, centered on the target residue and including the five spatially neighboring residues on each side. So, each residue is represented by a 220-component vector in the predictor based on the residue spatial sequence profile, and by an 11-component vector in the evolutionary rate-based predictor. For the predictor which combines residue sequence profile with evolutionary rate information, a 231-component vector is required for each amino acid residue.

A leave-one-out cross-validation strategy was employed to conduct the related subsequent experiments. In this strategy, one protein from our dataset was selected; then the SVMs were trained on the remaining proteins and the interaction sites of the selected protein were predicted. Here, $3 \times 69 = 207$ experiments are implemented, and the predictors were trained using all of the positive samples and the same number of negative samples extracted randomly from the training set in each experiment. Owing to the stochastic method used for selecting negative samples, the results could rarely be reproduced exactly for the same protein with another cross-validation run. Therefore, the entire cross-validation procedure was repeated five times, and the resulting performances were used to evaluate our method.

2.3. Evaluation measures of predictor performance

Generally speaking, prediction *accuracy*, whose value is the ratio of the number of correctly predicted residues to the total number of residues in experiment, is the best index for evaluating the performance of a predictor. However, only 34.8% of the data are interacting residues, which leads a rather unbalanced distribution of positive and negative samples. To assess our method objectively, another two indices are introduced in this paper, namely specificity and sensitivity [14,25].

The *specificity* is generally defined as the ratio of the number of matched residues between the predicted set and the actual set over the total number of predicted residues. The *sensitivity* is defined as the ratio of the number of matched interaction sites over the total number of the interaction sites in the observed set. Let TP be the number of true positives, i.e., residues predicted to be interface residues that actually are interface residues, and FP be the number of false positives, i.e., residues predicted to be interface residues that are in fact not interface residues. In addition, let TN be the number of true negatives, and FN the number of false negatives. Then the evaluation measures can be computed as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

$$\text{Correlation coefficient (CC)} = \frac{\text{TP} \times \text{TN} - \text{FP} \times FN}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{TP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

The *correlation coefficient* (CC) is a measure of how well the predicted class labels correlate with the actual class labels. Its range is from −1 to 1, where a *correlation coefficient* of 1 corresponds to perfect prediction and −1 to the worst possible prediction; a *correlation coefficient* of 0 corresponds to random guessing.

3. Results

3.1. Performance of three SVM predictors

The general performances of the three SVM predictors are shown in Table 1. The values of each measure are obtained by comparing the results from the five experiments, such that if a residue was predicted to be an interface residue no less than three times, it was taken as a positive prediction, and treated as a negative prediction otherwise. It can be seen that the difference in performance between the residue sequence profile-based predictor and the evolutionary rate-based predictor is very small. If judged by *accuracy* only, the evolutionary rate-based predictor seems to slightly outperform (by 2%) the sequence profile-based predictor. However, the *sensitivity* achieved by the sequence profile-based predictor is higher than that of the evolutionary rate-based predictor (7.7% better *sensitivity*), albeit with 1.7% lower *specificity* and an approximately equal *correlation coefficient*. The results indicate that the residue evolutionary rate approach can distinguish protein interaction sites from other positions on the protein surface, and its capability is almost identical to the residue sequence profile approach adopted by many previous studies to investigate protein–protein interaction.

It also can be found that the predictor whose feature vectors combined residue sequence profile with evolutionary rate outperforms the predictors based on either attribute alone. When both types of attributes are combined, the improvement in performance is impressive: at least 5% increase in *sensitivity*, 2% increases in *specificity* and *accuracy*, and 7% increase in *correlation coefficient*. These enhancements in all of the measures of performance used here indicate that the information contained within the residue sequence profile and the evolutionary rate may be complementary, and that exploiting this complementarity is helpful for predicting interaction sites.

Table 1
The overall performance of our experiments

	Sensitivity	Specificity	Accuracy	Correlation coefficient
Sequence profile	61.4%	45.8%	0.618	0.223
Evolutionary rate	53.7%	47.5%	0.637	0.220
Sequence profile + evolutionary rate	66.3%	49.7%	0.654	0.297

Table 2
The variances of different performance measures rooted from stochastic selection of negative samples across 69 proteins

	Sequence profile	Evolutionary rate	Sequence profile + evolutionary rate
Sensitivity	0.0070 ± 0.0013	0.0008 ± 0.0008	0.0039 ± 0.0050
Specificity	0.0019 ± 0.0032	0.0004 ± 0.0013	0.0008 ± 0.0011
Accuracy	0.0015 ± 0.0018	0.0002 ± 0.0003	0.0008 ± 0.0010
Correlation coefficient	0.0068 ± 0.0082	0.0006 ± 0.0009	0.0023 ± 0.0032

In each SVM training process, the negative sample was selected stochastically, so it is important to infer the influence of this randomness. This was studied by computing the variance of each performance measure through five repetitions across the 69 protein chains. The resulting means and standard deviations are shown in Table 2, and it can be seen that all the variances are close to zero. This result indicates that all the different SVMs converge to similar vectors, and it means our systems can predict interface residues through learning.

The detailed results of our experiments are depicted in Fig. 1. It can be seen that the predictor using both residue sequence profile and evolutionary rate as feature vectors outperforms that of the other predictors for all proteins, in almost all of the performance measures. The statistical analysis of each performance measure across all proteins also demonstrated this point, i.e., the combined attributes-based predictor achieves a higher mean and a lower standard deviation in almost all measures (Table 3). Unless otherwise noted, the following discussions in this section are based on the combined attributes-based predictor (Fig. 2). It can be seen that the *sensitivity* values were greater than 20% for all proteins, and at least 50% residues that were correctly classified for over 82% (57 of 69) of the proteins. The distribution of *specificity* values shows that this measure exceeds 50% in only 30 experiments and is less than the corresponding *sensitivity* values, indicating that there are relatively more false positives in our experiments. In the cases of 61 proteins, the prediction results can be regarded as credible if the cut-off of the *accuracy* is set at 0.5. Among the evaluation measures adopted here, the *correlation coefficient* values can best show how well our predictor worked. From Fig. 2, it can be found that for 96% of the proteins the *correlation coefficient* is greater than 0, which suggests that our predictor is indeed better than the random predictor [25].

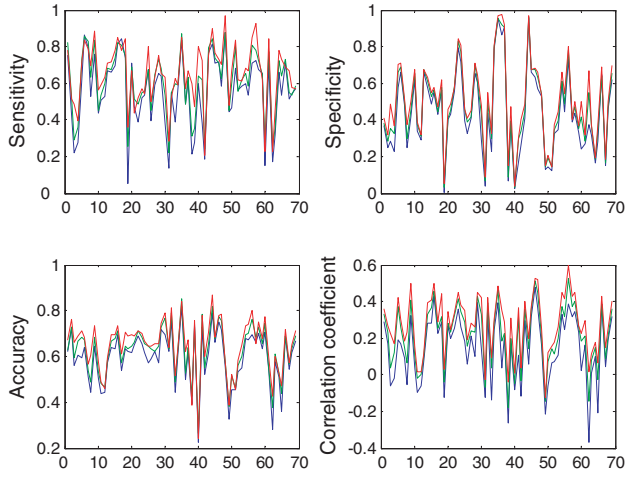

Fig. 1. The detailed performance measures across 69 proteins. Blue corresponds to sequence profile-based predictor; green denotes evolutionary rate-based predictor, and red denotes the combined attributes-based predictor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Location of interaction sites

To further illustrate the effectiveness of our approach, a test on protein complex 1BRL (PDB code) [26] was taken as an example. 1BRL is Luciferase, which is a class of enzymes that generate light in the visible spectrum, found in *luminous marine bacteria*. Its crystal structure has been determined to 2.4-Å resolution. It is an α–β heterodimer monooxygenase that catalyzes the oxidation of FMNH2 and a long-chain aliphatic aldehyde [27].

The prediction results are presented in Fig. 3, using the RasTop tool [28]. They showed that most interface residues and non-interface residues can be predicted correctly. Only 6.7% of the surface residues (24 false negative residues from a total of 365 surface residues) cannot be classified correctly into interface residues. Although 71 non-interface residues were predicted to be interface residues, we can remove most of them with the help of three-dimensional structure visualization of the target complex.

4. Discussions

This paper addresses the problem of distinguishing interface residues from other surface residues in heterocomplexes of known structure using SVMs. The results reported here demonstrate that residue sequence profile and evolutionary rate approaches can not only predict interface residues, but can also improve prediction performance by combining these two attributes. Interestingly, the prediction performances are nearly the same whichever of the two attributes was used as input vectors for the SVMs.

Table 3
The statistical analysis of predictor performance across 69 proteins

	Sensitivity	Specificity	Accuracy	Correlation coefficient
Sequence profile	55.6% ± 0.194	43.5% ± 0.225	0.593 ± 0.117	0.142 ± 0.185
Evolutionary rate	60.9% ± 0.172	47.4% ± 0.220	0.626 ± 0.112	0.218 ± 0.168
Sequence profile + evolutionary rate	65.0% ± 0.167	50.4% ± 0.220	0.650 ± 0.116	0.274 ± 0.161

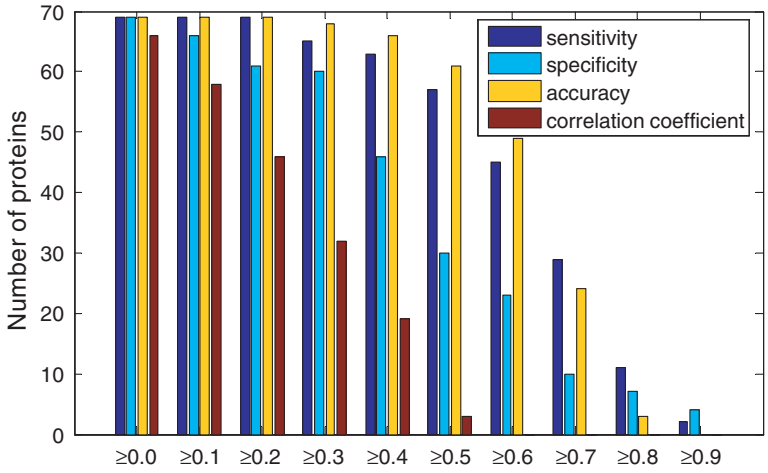

Fig. 2. The distributions of prediction performance measure values of the combined attributes-based predictor for 69 proteins.

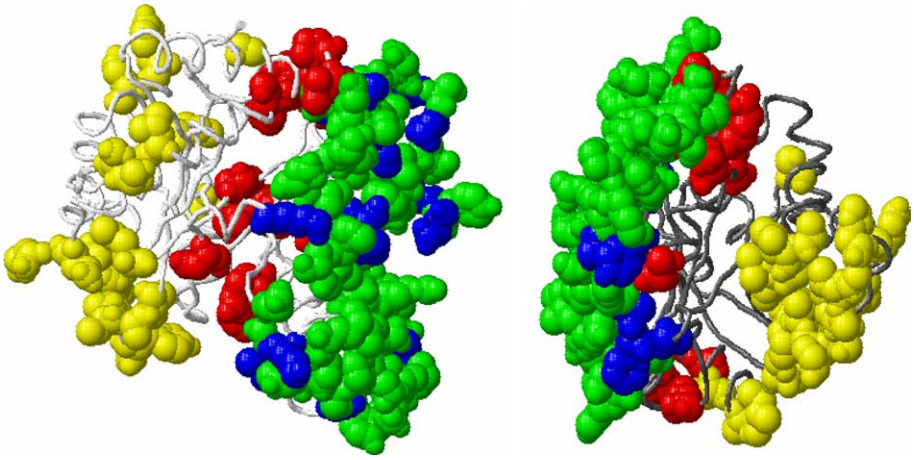

Fig. 3. Visualization of prediction results on heterocomplex PDB:1BRL. The 3D structure of the complex is shown by a smooth spline between consecutive alpha carbon positions; white and black represent chains A and B, respectively. The residues related to the prediction are displayed as spheres and the corresponding colors are coded as follows: green denotes true positive predictions (TP); blue denotes the missing interface residues in the predictor (FN); red and yellow denote false positive predictions (FP), of which the yellow residues can be excluded by visualization of the complex's 3D structure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

There are several methods using neural network or SVM to identify interface residues based on spatially or sequentially neighboring residue profiles [10–16]. We repeated the training processes using back-propagation neural networks [29] on the same dataset, and the performances demonstrated the effectiveness of the proposed SVM algorithm in tackling this problem (Fig. 4). Some previous studies [13–15] have adopted an SVM algorithm to study protein interaction sites from primary structure. A direct comparison with these studies is difficult due to the differences in choice of dataset and definitions of surface or interface residue. But it is clear that predicting protein–protein interaction sites from sequentially neighboring residues is harder than from spatially neighboring sequences in the absence of structure information; this is important because there is biological importance in revealing the function of proteins whose structure are known.

A relatively high false positive ratio in protein–protein interaction sites prediction is a troublesome problem. Some investigators reduce the false positive ratio by eliminating isolated raw positive predictions [11,15]. For structure-known proteins, we can exclude false positive predictions by considering their three-dimensional structure. On the other hand, these false positive predictions might comprise other functionally important sites which do not correlate directly with protein–protein interactions in our selected complexes, but rather imply potential interactions between the target protein and other proteins in a specific environment.

The results obtained in this paper show that our proposed method is a promising approach for studying protein–protein interaction. The protein–protein interaction residues are more likely to remain unchanged during evolution. Though this study only includes 69 protein chains, as a methodology based on evolutionary conservation, the predictor can be well generalized for new structure-known proteins. Predictions generated here should facilitate experimental investigators to validate the roles of specific residues in protein complexes. Incorporation of our approach with physicochemical or geometric properties and other attributes of interaction regions will yield progress for studying protein–protein interactions.

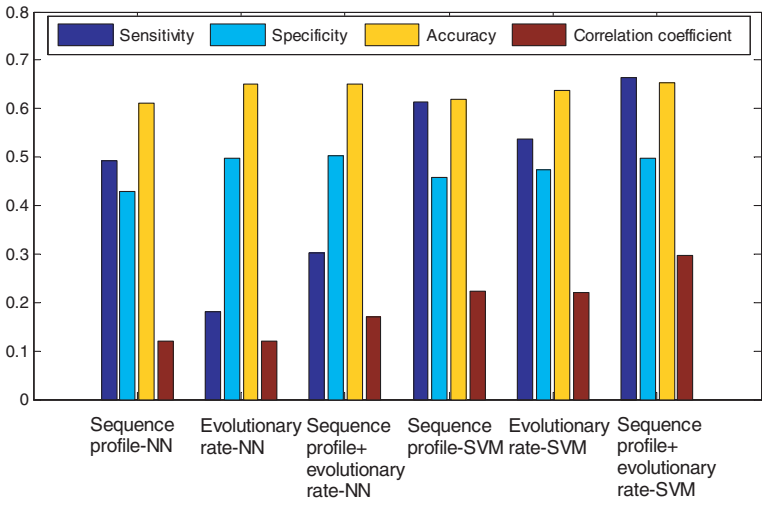

Fig. 4. Comparisons of the predictor performance using neural networks (NN) with SVM algorithm. Sequence profile-NN denotes the NN predictor using sequence profile as input vectors, and the similar labels are employed by the other five predictors.

Acknowledgments: This work was supported by the National Science Foundation of China (Nos. 60472111, 30570368 and 60405002). The work described in this paper was partially supported by three grants from the Hong Kong Special Administrative Region, China: RGC Project Nos. CUHK 4170/04E, 4205/04E and UGC Project No. AoE/E-01/99.

References

[1] Alberts, B.D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1989) Molecular Biology of the Cell, 2nd edn, Garland, New York.
[2] Chelliah, V., Chen, L., Blundell, T.L. and Lovell, S.C. (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J. Mol. Biol. 342, 1487–1504.
[3] Kini, R.M. and Evans, H.J. (1996) Prediction of potential protein–protein interaction sites from amino acid sequence identification of a fibrin polymerization site. FEBS Lett. 385, 81–86.
[4] Janin, J. (1997) Specific vs. non-specific contacts in protein crystals. Nat. Struct. Biol. 4, 973–974.
[5] Thorn, K.S. and Bogan, A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17, 284–285.
[6] Pazos, F. and Valencia, A. (2002) In silico two hybrid system for the selection of physically interacting protein pairs. Proteins 47, 219–227.
[7] Gallet, X., Charloteaux, B., Thomas, A. and Brasseur, R. (2000) A fast method to predict protein interaction sites from sequences. J. Mol. Biol. 302, 917–926.
[8] Jones, S. and Thornton, J.M. (1997) Analysis of protein–protein interaction sites using surface patches. J. Mol. Biol. 272, 121–132.
[9] Jones, S. and Thornton, J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. J. Mol. Biol. 272, 133–143.
[10] Zhou, H. and Shan, Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44, 336–343.
[11] Ofran, Y. and Rost, B. (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett. 544, 236–239.
[12] Fariselli, P., Pazos, F., Valencia, A. and Casadia, R. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur. J. Biochem. 269, 1356–1361.
[13] Yan, C., Dobbs, D. and Honavar, V. (2004) A two-stage classifier for the identification of protein–protein interface residues. Bioinformatics 20, i371–i378.
[14] Yan, C., Dobbs, D. and Honavar, V. (2003) Identification of residues involved in protein–protein interaction from amino acid sequence – a support vector machine approach in: Intelligent Systems Design and Applications (Abraham, A., Franke, K. and Köppen, M., Eds.), pp. 53–62, Springer, Berlin, Germany.
[15] Res, I., Mihalek, I. and Lichtarge, O. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics 21, 2496–2501.
[16] Koike, A. and Takagi, T. (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng. Des. Sel. 17, 165–173.
[17] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
[18] Rost, B. and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20, 216–226.
[19] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
[20] Hua, S. and Sun, Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.
[21] Dodge, C., Schneider, R. and Sander, C. (1998) The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res. 26, 313–315.
[22] Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. and Ben-Tal, N. (2005) The ConSurf-HSSP Database: the mapping of evolutionary conservation among homologs onto PDB structures. PROTEINS: Struct. Function Bioinformatics 58, 610–617.
[23] Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18, s71–s77.
[24] Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19, 163–164.
[25] Baldi, P., Brunak, S., Chauvin, Y. and Andersen, C.A.F. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.
[26] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235–242.
[27] Fisher, A.J., Raushel, F.M., Baldwin, T.O. and Rayment, I. (1995) Three-dimensional structure of bacterial luciferase from Vibrio harveyi at 2.4 Å resolution. Biochemistry 34 (20), 6581–6586.
[28] RasTop-Molecular Visualization Software. Available from: <http://www.geneinfinity.org/rastop>.
[29] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning representations by back-propagating errors. Nature 323, 533–536.