Rapid Communication

Prediction of protein binding sites using physical and chemical descriptors and the support vector machine regression method*

Sun Zhong-Hua(孙重华)[a)b)] and Jiang Fan(江 凡)[a)†]

[a)] *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*

[b)] *Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*

(Received 25 June 2010; revised manuscript received 7 July 2010)

In this paper a new continuous variable called core-ratio is defined to describe the probability for a residue to be in a binding site, thereby replacing the previous binary description of the interface residue using 0 and 1. So we can use the support vector machine regression method to fit the core-ratio value and predict the protein binding sites. We also design a new group of physical and chemical descriptors to characterize the binding sites. The new descriptors are more effective, with an averaging procedure used. Our test shows that much better prediction results can be obtained by the support vector regression (SVR) method than by the support vector classification method.

Keywords: protein binding site, support vector machine regression, cross-validation, neighbour residue

PACC: 0500, 3620, 8200, 8700

1. Introduction

Computational methods of predicting the protein binding sites can reveal the essential characteristics of protein interaction sites, and are very important for studying the mechanism of protein interactions.[1] Moreover, the accurate prediction of protein binding sites has many useful applications in protein docking[1,2] and in structural model building of protein–protein complexes.

To analyse the surface characteristics of protein binding sites, the surface residues of a protein molecule are divided into two classes.[1] One is the surface residues which are in contact with another molecule in a protein–protein complex and defined as the interface residues, and the other is the surface residues which are defined as the non-interface residues. Experimental results show that the interface residues and the non-interface residues have significant differences in characteristic. This difference forms the basis of designing computational methods to predict protein binding sites.

Previous studies have shown that the interface residues and the non-interface residues are different mainly in the following aspects. First, lots of work shows that the interface residues are conserved more than the non-interface residues.[3−5] Most of the site prediction methods include the sequence conservation as one of the descriptors for characterizing the interface, and moreover, a few methods[5,6] use only the sequence conservation and can still obtain somewhat accurate results. Second, Chen and others found that the solvent accessible surface of interface residues was greater than the non-interface residues when they were calculated in unbound state.[7,8] So this descriptor is often used for describing the residues of a protein binding site. Third, in Chen's study,[7] it is found that the most preferable interface residues include leucine, isoleucine, phenylalanine, tyrosine and valine. Apparently they are all hydrophobic amino acids; therefore, many studies include the hydrophobic indexes in various formats in the description of the interface residues.

In our previous work, we designed a prediction method to classify the interface residues and the non-interface residues using the support vector classification (SVC) algorithm.[9] Further studies showed that this method had some limitations. A binding site can be divided into the core region and the rim

*Project supported by the National Natural Science Foundation of China (Grant Nos. 10674172 and 10874229).

†Corresponding author. E-mail: fjiang@aphy.iphy.ac.cn

© 2010 Chinese Physical Society and IOP Publishing Ltd

region.[10,11] And the residues in the respective regions are somewhat different: the rim interface residues are more likely to resemble the non-interface residues and the core interface residues are more likely to resemble the interior of the protein, because they are often completely buried in the complex. This is consistent with the fact that the core interface residues are more conserved in sequence, more hydrophobic, and more buried as mentioned in the above. For this reason, we design a nearly continuous variable to better describe the position of a surface residue being in an interface, near the rim of an interface, or out of an interface. This is a better description than the binary description as we have done previously. With this continuous description using a variable, named the core-ratio, we can now use the support vector regression (SVR) algorithm to learn and fit the variable core-ratio using a feature vector of descriptors. The trained model obtained can then be used to predict the residues in a protein binding site.

2. Methods and materials

2.1. SVR theory and lib-SVM software

Support vector machine (SVM) is a machine learning algorithm proposed by Vapnik[12] in 1995. SVR is an application of the SVM algorithm to solve the regression problem. SVR algorithm uses the kernel function to do a functional mapping so that a nonlinear input space can be mapped into a high dimensional linear feature space. Consequently the linear statistical learning machine can be applied to this high dimensional kernel space to achieve a binary classification and hence solve the nonlinear classification problem. In the present work, the radial basis kernel function is used, which is also widely used in other bioinformatics problems.

In the present work, we use a lib-SVM software which was distributed freely by Fan *et al.*[13] The Lib-SVM programs provide linear, polynomial, radial basis function (RBF), and S-shaped kernel functions, and C-SVC, n-SVC, e-SVR, n-SVR, one-class-SVM optimization algorithms. We choose RBF kernel and e-SVR algorithm.

2.2. Definition of the core-ratio.of surface residues

The core-ratio of a surface residue is defined as the number of the interface residues divided by the number of all surface residues within a given distance around the surface residue whose core-ratio is to be calculated. The position of a surface residue is determined by its C_β atom position. In this work, the radius is 15 Å (1 Å=0.1 nm), that is, $C_\beta - C_\beta$ distance between two residues should be less than 15 Å. The core-ratio value ranges from 0 to 1. Outside the bind site the core-ratio is zero valued. A residue with a core-ratio being close to 1.0 suggests that around this residue, almost all other residues are interface residues. Thus this residue is located in the centre or core region of the binding site.

2.3. Descriptors of a surface residue

Besides the three significant descriptors, we include other physical and chemical descriptors to fully characterize a surface residue. As a result, there are sixty descriptors for the feature vector of a residue. They are divided into four classes.

The first class of descriptors is for describing the characteristics of a given residue, which is also called the centre residue, as some features in this class are also related to residues flanking it in the peptide sequence. This class consists of eight items, including the number of heavy atoms, the maximum number of potential hydrogen bonds, the net charge, the hydrophobic index, the sequence conservation, the secondary structure, the relative solvent accessible surface, and the polar fraction of the side chain surface area for the given amino acid type of the centre residue. Of the eight items, the first four items are only related to the centre residue itself. The sequence conservation is based on the primary sequence segment flanking the center residue and calculated by using psi-Blast.

The second structure, the relative solvent accessible surface, and the polar fraction of the side chain surface area are based on the tertiary structure of the protein. They are calculated by the DSSP[14] and CCP4 SURFACE programs.[15,16] The second class of descriptors is for describing the average characteristics of the neighbouring surface residues around the centre residue within a radius of 15 Å. The distance between two residues is defined as their $C_\beta - C_\beta$ distance. This class contains ten items, among them eight are the same as the first class but with averaged values within the 15 Å radius, and the other two include the average sequence distance and the average spatial distance of the neighbouring residues relative to the centre residue.

The third class of descriptors is for describing the residue composition and the atom composition of the neighbouring surface residues within a radius of 15 Å, with the centre residue included. The residue composition is the number of each type of amino acid divided by the total number of the neighbouring residues. There are twenty types of amino acids, so the descriptor of the residue composition has twenty items. Similarly, we calculate the atom composition for the neighbouring atoms according to atom types as opposed to amino acid types for the residue composition. We use the DeLisi atom types, which have eighteen types.[17]

The fourth class of descriptors is for describing the relative size and shape of the surface and volume around the centre residue by using the number of the neighbouring residues and the number of the neighbouring atoms. This class has four items, including the number of the neighbouring surface residues, the number of all neighbouring residues including the interior residues, the number of the neighbouring surface atoms, and the number of all neighbouring atoms including the interior atoms. These descriptors can reflect the local size and the shape of a centre residue in a protein structure.

2.4. Data set

In our test we use 977 complexes of heterologous protein partners. The training set includes 927 complexes and the test set includes 50 complexes. These complexes are chosen from the Protein Data Bank (PDB) database. The complex structure must be obtained by the x-ray method, and the resolution must be greater than 3.0 Å. The complex must be composed of heterologous proteins. The protein chain must be longer than forty residues. In each complex, each partner can have only one protein chain. The homology of any two chains in all complexes in the dataset must be less than 30 per cent.

2.5. Evaluation of test results

First we use the n-fold cross validation to evaluate the effectiveness of our training. In the n-fold cross-validation, the original sample is randomly partitioned into n subsamples. Of the n subsamples, a single subsample is retained as the validation data for testing the trained model, and the remaining $(n-1)$ subsamples are used as the training data. The cross-validation process is then repeated n times (the n folds), with each of the n subsamples used exactly once as the validation data. The n results from the n folds can then be combined to produce a single estimation of the cross-validation. The estimations in the SVR method are the cross validation correlation coefficient (CC) and the cross validation mean squared error (MSE):

$$\mathrm{CC} = \sqrt{\frac{(n*\sum_{i=1}^{n}(y'_i*y_i)-\sum_{i=1}^{n}(y'_i)*\sum_{i=1}^{n}(y'_i))^2}{(n*\sum_{i=1}^{n}(y'^2_1)-\sum_{i=1}^{n}(y'_i)*\sum_{i=1}^{n}(y'_i))*(n*\sum_{i=1}^{n}(y_i^2)-\sum_{i=1}^{n}(y_i)*\sum_{i=1}^{n}(y_i))}},$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2,$$

where y_i is the experimental value to be fitted by the regression method and the y'_i is the fitted value.

We calculate the accuracy and coverage[1] of the predicted results for the test set. Accuracy (ACC)=$T_\mathrm{p}/(T_\mathrm{p}+F_\mathrm{p})$, where T_p is the number of true positive (correct) predictions, and F_p is the number of false positive (wrong) predictions. Coverage (COV)=$T_\mathrm{p}/(T_\mathrm{p}+F_\mathrm{n})$, where F_n is the number of false negative (wrong) predictions.

3. Results and discussion

3.1. Core-ratio distribution

The value of the core-ratio is between zero and one. A higher core-ratio value means that the density of interface residue around the centre residue is greater. This indicates that the residue is located in the centre of the interface. A core-ratio value equal to zero means that there is no interface residue around it, so this is a non-interface residue. We divide the [0, 1] interval into ten subintervals and count the number of residues whose core-ratios are in each of the subintervals. The resulting distribution of the core-ratio values is shown in Fig. 1. It can be seen that the core-ratio is a good continuous variable for describing the surface residue position relative to the centre of the interface. So it is appropriate that we use the core-ratio value as the fitting variable for the SVR regression to predict the interface residues in the protein binding sites.

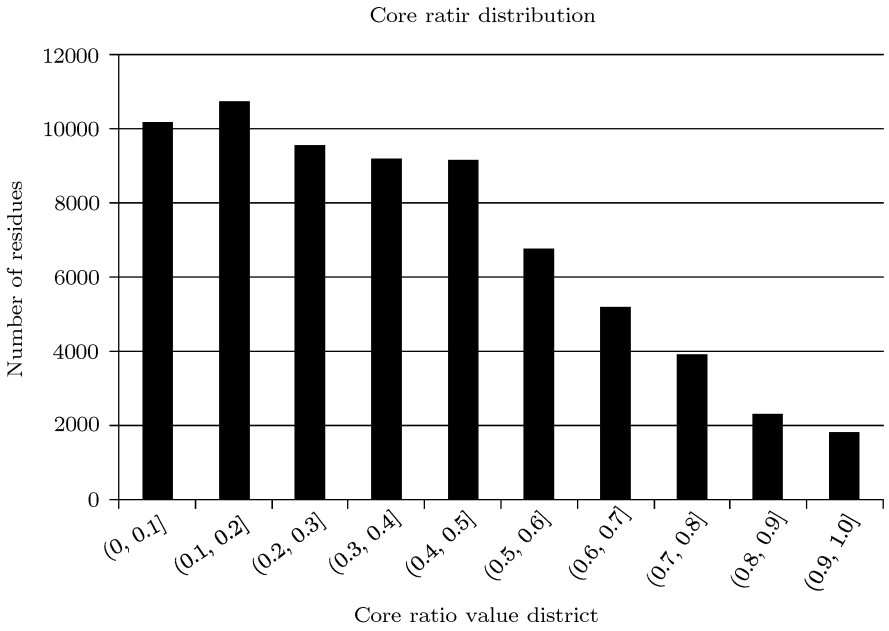


Fig. 1. Histogram of core-ratio distribution, where the interval [0, 1] of core-ratio value is divided into ten subintervals and the number of residues whose core-ratios are in each subinterval are counted. This figure shows that there are fewer residues when the core-ratio is approaching to 1. Such a distribution of the core-ratio is a good description of the physical distribution of the interface residues and it is almost continuous as expected.

3.2. Residue composition according to three classes surface residues

We divide surface residues into three classes according to their core-ratio values. The core interface residues should have their core-ratios no less than 0.5. The rim interface residues should be greater than 0 and less than 0.5. If a residues has a core-ratio equal to 0, it is referred to as a non-interface residue. The compositions of the twenty amino acids for the three classes of surface residues are shown in Fig. 2. From the figure we find that the eight amino acids have strong and clear patterns. Five of them are hydrophobic amino acids and they are phenylalanine, leucine, tyrosine, isoleucine and valine.

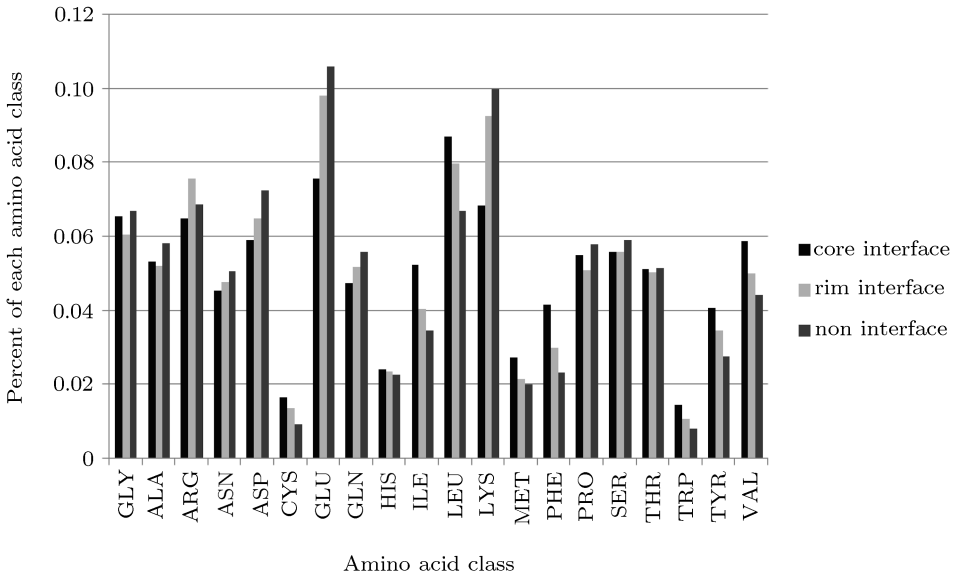


Fig. 2. Residue compositions in three classes surface residues where the compositions of twenty amino acids in the three classes of surface residues are calculated and five types of hydrophobic amino acids are found to be more in the core interface residues than in the non-interface residues, these being phenylalanine, leucine, tyrosine, isoleucine and valine. In contrast, three types of charged amino acids prefer to be the non-interface residues, these being lysine, glutamic acid and aspartic acid.

These amino acids each strongly prefer being an interface residue to being one of interface residues in other two classes, in particular, the non-interface residues. That is, the non-interface residues are much less hydrophobic than the core–interface residues. On the other hand, lysine, glutamic acid and aspartic acid prefer to be in the class of non-interface residues, all of which have the charged side chains. The other twelve amino acids do not have a clear trend, so they are less effective in distinguishing the residues between the interface and the non-interface regions.

3.3. Comparison of the accuracy and coverage of the prediction results

We use the SVC method and the SVR method to predict the same test set of fifty protein complexes and compare their results. When we use the SVR method, the cut-off value used to distinguish the positive prediction from the negative prediction is 0.6. If the predicted core-ratio value of a residue is more than 0.6, it will be predicted as an interface residue, that is, a positive prediction. Otherwise the residue will be predicted as a non-interface residue, that is, a negative prediction. By comparing the predicted results with the definitions based on the experimental complex structures, we can judge whether the predicted results are true. Hence we can calculate the accuracy and the coverage of the total predicted results for the test set. The results are shown in Table 1. We can find that under the same coverage the SVR method achieves a much better accuracy of 0.651 than the SVC method.

Table 1. Comparison of ACC and COV results of the SVC method and the SVR method.

training method	ACC	COV
SVC method	0.503	0.283
SVR method	0.651	0.277

3.4. Effects of four classes of descriptors on cross validation error of the training set

The descriptors of surface residues are divided into four classes in our method. We test the effect of each class of descriptors on the training models. First, we make a training set using only the eight descriptors in the first class and perform the cross validation. Then we add the other three classes of descriptors into the training set one by one. These four training tests include the same residues in the training set, and the cross validation results are shown in Table 2. We can see that when we add each additional class of descriptors, the CV_CC (cross-validation of CC which is defined in the above equation) will increase while the CV_MSE (cross-validation of MSE which is defined in the above equation) will decrease. It means the new combined class of descriptors is more effective than the previous class. Moreover, when we add the second class of descriptors, the cross validation results are improved by the largest amount. It indicates that the characteristics of the environment around the centre residue are very important for describing the centre residue in the interface.

Table 2. Effects of four classes of descriptors on the cross validation error of the training set.

class of descriptors	CV_CC	CV_MSE
8 item-descriptors for the center residue	0.337	0.0700
(8+10) item-descriptors adding the neighboring residues	0.561	0.0542
(8+10+38) item-descriptors adding residue and atom compositions	0.583	0.0521
(8+10+38+4) item-descriptors adding size and shape	0.589	0.0515

3.5. Discussion improved by the largest amount

In our novel method of using the SVR method to predict the binding site residues, we first define the core-ratio variable to describe the distribution of the interface residues on a protein surface. It is shown that the continuous core-ratio is more useful than the binary descriptor of the interface residue using 0 and 1. As a result, we can use the core-ratio as the fitting value for the SVR method to predict the protein binding sites instead of using the SVC method, as was done previously. Table 1 shows that much better prediction results have been obtained by the SVR method than

by the SVC method.

Previously, when we tested the SVC method on the crystal structures of the complexes using either the bound components or the unbound components, very different results were obtained. The accuracy of the predictions using the unbound components was much worse than using the bound components, which we have attributed to the conformational changes caused by the complex formation. To alleviate this effect of the conformational changes especially in the side chains, we designed an averaging method to smooth out the effect of conformational change of the individual side chains.

Furthermore, previous studies, ours' and others', have shown that the environment .of a residue is important for the prediction of protein binding sites. In the SVC method, we used a fixed number of spatial neighbouring residues and sequential neighbouring residues of a centre residue, similar to other studies. It was found that the local packing was different from complex to complexe, so this description is sometimes not good enough. So, in the present study, in the SVR method we use the average values of physical and chemical descriptors to describe the environment of the neighbouring residues around a centre residue within a distance radius of 15 Å. The new environment descriptors in the SVR method can also reduce the bad effect induced by the conformational changes compared with those in the SVC method, as discussed above. As a result, the total number of descriptors reduces from 122 items for the SVC method down to 60 items for the SVR method, so the computational speed is greatly enhanced and over-fitting due to lack of training data is further avoided. From the prediction results we can see that the new descriptors have better performances. It should be emphasized that the method of averaging the properties of the neighbouring residues and using the environment characteristics has not been tested by other investigators.

In our study, we find that the training data set will affect the prediction results somewhat. In our method we incorporate all kinds of biological functional protein interactions into one training set. The resulting model can predict most of the protein binding sites effectively. But for a certain subset of protein complexes, the prediction results may be not best. In future work, we will classify the protein interactions according to either biological functions or other properties, and subdivide the training set for each class of interactions, thereby improving the prediction accuracy further and conducting to the study of the difference between different classes of protein interactions.

Finally, our method of using the SVR algorithm can be extended to the prediction of binding sites for other macromolecules such as RNA and DNA.[18] Furthermore, more statistical properties can be designed to incorporate the principles of protein structures and molecular interactions[19−22] into the descriptors of molecular binding sites such as folding codes, fractal dimensions and shape factor.

References

[1] Zhou H X and Qin S B 2007 *Bioinformatics* **23** 22032209
[2] Smith J R and Sternberg M J 2002 *Curr. Opin. Struct. Biol.* **12** 28
[3] Hu Z, Ma B, Wolfson H and Nussinov R 2000 *Proteins* **39** 331
[4] Ma B, Elkayam T, Wolfson H and Nussinov 2003 *Proc. Natl Acad. Sci.* USA **100** 5772
[5] Armon A, Graur Dan and Ben-Tal N 2001 *J. Mol. Biol.* **307** 447
[6] de Vries S J, van Dijk A D J and Bovin A M J J 2006 *Proteins* **63** 479
[7] Chen H and Zhou H X 2005 *Proteins* **61** 21
[8] Janin J, Miller S and Chothia C 1988 *J. Mol. Biol.* **204** 155
[9] Li N, Sun Z and Jiang F B M C 2008 *Bioinformatics* **9** 553
[10] Chakrabarti P and Janin J 2002 *Proteins* **47** 334
[11] Bahadur R P, Chakrabarti P, Rodier F and Janin J 2003 *Proteins* **53** 708
[12] Vapnik V 1995 *The Nature of Statistical Learning Theory* (New York: Springer)
[13] Fan R E, Chen P H and Lin C J 2005 *Journal of Machine Learning Research* **6** 1889
[14] Kabsch W and Sandor C 1983 *Biopolymers* Dec **22** 2577
[15] Collaborative Computational Project Number 4. 1994 *Acta Crystallogr* D **50** 760
[16] Lee B and Richards F M 1971 *J. Mol. Biol.* **14** 379
[17] Zhang C, Vasmatzis G, Cornette J L and DeLisi C 1997 *J. Mol. Biol.* **267** 707
[18] Gao L F, Liu X and Guan S 2008 *Chin. Phys.* B **17** 4396
[19] Liu J F 2009 *Chin. Phys.* B **18** 2615
[20] Jiang F and Li N 2007 *Chin. Phys.* **16** 392
[21] Xiao Y and Yao K L 1994 *Chin. Phys.* **3** 788
[22] Wang X H, Shen Y and Zhang L X 2009 *Chin. Phys.* B **18** 1684