Neural Comput & Applic (2007) 16:481–490
DOI 10.1007/s00521-006-0076-4

ORIGINAL ARTICLE

Predicting the distance between antibody's interface residue and antigen to recognize antigen types by support vector machine

Yong Shi · Xinyang Zhang · Jia Wan · Yong Wang · Wei Yin · Zhiwei Cao · Yajun Guo

Received: 14 November 2005 / Accepted: 3 November 2006 / Published online: 25 November 2006

Abstract In this paper, a machine learning approach, known as support vector machine (SVM) is employed to predict the distance between antibody's interface residue and antigen in antigen–antibody complex. The heavy chains, light chains and the corresponding antigens of 37 antibodies are extracted from the antibody–antigen complexes in protein data bank. According to different distance ranges, sequence patch sizes and antigen classes, a number of computational experiments are conducted to describe the distance between antibody's interface residue and antigen with antibody sequence information. The high prediction accuracy of both self-consistent and cross-validation tests indicates that the sequential discovered information from antibody structure characterizes much in predicting the distance between antibody's interface residue and antigen. Furthermore, the antigen class is predicted from residue composition information that belongs to different distance range by SVM, which shows some potential significance.

Keywords Bioinformatics · Protein · Protein data bank · Antibody–antigen complexes · Support vector machine · Cross-validation

Y. Shi (✉) · X. Zhang · W. Yin
Research Center on Data Technology
and Knowledge Economy, Graduate University
of Chinese Academy of Sciences,
Beijing 100080, China
e-mail: yshi@gucas.ac.cn

J. Wan
Institute of Biophysics, Chinese Academy of Sciences,
Beijing 100101, China

Y. Wang
Department of Electrical Engineering and Electronics,
Osaka Sangyo University, Daito,
Osaka 574-8530, Japan

Z. Cao
Shanghai Center for Bioinformation Technology,
Shanghai 200235, China

Y. Guo
Shanghai International Cancer Institute of China,
Shanghai 200433, China

1 Introduction

The explosive growth in biotechnology combined with major advances in information technology has the potential to radically transform immunology in the post genomics era [1]. Indeed, the computational immunology aims to provide tools for extraction, comparison, analysis and interpretation on not only the vast quantities of existing data, but also the newly accumulated data with relevance to immunology.

Antibodies are similar immunoglobulins in sequence and structure. According to the ratio of different amino acids in a given position among different antibodies, each antibody can be divided into constant and variable regions. Antibody binds its corresponding antigen by complementarity determining regions (CDRs), which are also known as hypervariable (HV) loops because of their dramatic variability as relative to others in the variable regions. Much effort has focused on characters of antibody structure, antibody–antigen binding site, the affinity and specificity of the antibody [2–8]. A good fit of the two binding surfaces in antibody–antigen complex is important for high affinity [9]. The circumstantialities of antibody–antigen interaction

surface can be observed by the distance between antibody's interface residue and antigen. It will help us to understand position of each interface residue relative to antigen in 3D, which connects affinity of antibody–antigen interaction.

In bioinformatics field, there are many studies of protein–protein interaction by different machine learning approaches [10–13]. Fariselli and Casadio [14] and Fariselli et al. [15] used neural network methods to study protein–protein contact sites based on chemicophysical and evolutionary information. The reported precision is about 73%. Similarly, Ofran and Rost [16, 17] predicted protein–protein interaction sites from sequence information by using neural network. Besides neural network, support vector machine (SVM) has been applied the field [18, 19]. In these studies, the protein–protein complexes were divided into six classes in order to improve the accuracy. Each class was analyzed and predicted using sequence information. Among these six classes, antibody–antigen complex was studied as a specific one. The sensitivity is 82.3%, specificity is 81.0% and correlation coefficient is 0.43. However, the characters of residue, such as distance range of antibody and antigen, need to be further investigated.

To promote the direction of research, this paper proposes a method of using machine learning to predict the distance between antibody's interface residue and antigen by characters of residue sequence in antibody structures as a pilot study to discover the correlation between antibody–antigen interaction residue and antibody–antigen interaction surface.

In this research, the proposed framework of the methods for predicting the distance between antibody's interface residue and antigen includes three stages:

Stage The surface residues and the interior residues are distinguished from protein structure information. To find surface residues, relative solvent accessibility is employed. Relative solvent accessibility is an important character in protein structures, which has been extensively studied [20–24]. The commonly known program DSSP [25] is employed to compute accessible surface area (ASA). Threshold values are used to determine the binary categories.

In our study, a kind of complex structure is designated as the basic standard in the paper. It is extracted from the antibody–antigen complexes, which is selected from protein data bank (PDB) [26]. The complex structure is composed by heavy chains, light chains and the corresponding antigens. It is then modified by fixing the missing residues in heavy chain or in light chain sequences. After removing the redundant information, 37 basic structures are extracted. The total of 5,253 surface residues is then selected from these 37 complex structures to form data set.

Stage To distinguish interface residue. Jones and Thornton [27] reported that if the ASA calculated from one surface residue is 1 Å less than the value calculated from monomer, then this surface residue is regarded as an interface residue. Fariselli et al. [15] also indicated that if the Cα distance of two neighboring surface residues is less then 12 Å, these two surface residues are regarded as the contacted residues. The latter standard is adopted by other researchers for its simplicity and easy implement. Surface residues are defined as interface residues if surface residues contact other residues in different structure. In this paper, if the calculated distance is less than 12 Å, it is defined that residue contacts antigen. The total of 668 interface residues is then selected from these 5,253 surface residues.

Stage To predict the distance range between antibody's interface residue and antigen. In order to study the distance-range, the distance from Cα in interface residues to the nearest atom in antigen is calculated. If the calculated distance is less than the threshold value, it is defined that the distance between interface residue and antigen belong to this range. The distance 8, 10 and 12 Å are selected as the threshold values to cope with range. The prediction accuracies derived from these distance ranges are compared and analyzed. There are 329, 508, 668 interface residues belong to distance ranges 8, 10, 12 Å, respectively.

With the preparation of the data set, a machine learning framework in this paper can be considered as a two-step process:

Step To extract feature vector from the sequence information. In this paper, two coding methods are proposed for handling the sequence feature, amino acid composition, etc. A scheme is also proposed to extract feature vector from information of sequence neighbors.

Step To utilize experimental scheme to mine the useful information by using support vector machine (SVM). The details of the dataset, methodology and evaluation measures are reported in Sect. 2.

In this study, two factors, the sequence patch size and the antigen classes are mainly explored to exam how they affect the prediction accuracy.

In our experiments, five sequence patch sizes are chosen. They are 1, 2, 3, 4 and 5, respectively. The prediction accuracies of these sequences are also investigated as follows.

It is also concerned that how antigen classes affect the antibody–antigen interaction surface. Different types of antibody combining sites have been studied

[2]. These are cavity or pocket (hapten), groove (peptide, DNA, carbohydrate) and planar (protein). In our research, antigens are divided into two classes (protein, non-protein) for study. Meanwhile the antigen classification influence is studied by comparing the accuracy with classification to the accuracy without classification. Another kind of antigen class (protein and non-protein) is not subject to classification.

There are totally 45 samples produced according to different distance ranges, sequence patch sizes and antigen classes in the data set. The samples are trained and tested using SVM for prediction the distance range between antibody's interface residue and antigen.

Based on the prediction results of interface residues that belong to different distance ranges, another experiment is designed to predict the class of antigen because different types antibody–antigen interface have different surface characters [2]. In our study, the number of each kind of interface residues belong to different distance range is calculated as a basis of composing the attributes of interface residue.

There are totally three samples are produced according to three distance ranges in our data set. The SVM is used again to train and test so that the functional class from antibody structures data can be identified.

These numerical results are shown and discussed in Sect. 3, while the conclusion is given in Sect. 4.

2 Materials and methods

2.1 Collection of complex structure

The structural data of antibody–antigen complex exists in relevant biological literature and databases. In this paper, all the antigen–antibody compound files are selected from the PDB file library. The heavy chain, light chain and corresponding antigen from the complex are collectively defined as the basic complex structure, which is the start point of our research.

After testing the selected basic structure, some missing residues in the heavy chain and light chain of these structures are detected. We fill these missing residues using [HyperChem 5.1 for Windows (Hypercube, FL, USA)].

If the heavy chains and the light chains of antibody in two complex structures are exactly the same, one of the structures is defined redundant. After examination of all the basic structures, 37 non-redundant complex structures are extracted.

In the 37 complex structures obtained, 3 of the antigens are nucleic acid, 4 of the antigens are heterocomplex, and the rest are proteins. To simplify the problem, the antibody structures are further divided into two classes according to whether the antigen is protein or not. Meanwhile the antigen classification influence is studied by comparing the accuracy with classification to the accuracy without classification. Another kind of antigen class (protein and non-protein) is not subject to classification.

2.2 Antibody surface residue and antibody interface residues

The residues in the antibody structure can be divided into two groups: (1) residues embedded in the structure; (2) residues in the surface of the structure. It is commonly accepted that the surface residues in the antibody structure plays a vital role in interaction between antibody and antigen.

The accessible surface area method is utilized in the recognition of surface residues. The ASA is calculated using the DSSP program. If the ASA in the chain structure is 25% larger than the value calculated by the residue alone, this residue can be regarded to surface residue of the antibody structure.

After computation, 5,253 surface residues are recognized from the 37 non-redundant complex structures. Among them, 4,139 surface residues' antigens are of the protein class, and the other 1,114 surface residues' antigens are of the non-protein class.

Antibody function is accomplished by combination of antibody–antigen interaction surface. Interface residues in antibody play an important role in antibody–antigen interaction. There are many methods to identify interface residues from protein–protein complex. Fariselli et al. [15] indicated that two surface residues are regarded as interface residues, if distance between Cα atom of one surface residue in one protein and Cα atom of one surface residue in another protein is less then 12 Å.

In this paper, the coordinate of Cα atom in residue is taken as the coordinate of this residue. Distances between one surface residue in the antibody structure in one complex structure and every atom in antigen in this complex structure are calculated to identify whether this surface residue is interface residue or not. If the distance is less than 12 Å the surface residue is considered to contact with the antigen, and it is identified as interface residue.

After computation, 668 interface residues are recognized from 5,253 surface residues. Among them, 532 interface residues' antigens are of the protein class, and the other 136 interface residues' antigens are of the non-protein class.

For the research of the distance between antibodies interface residue and antigen, different ranges are used here. Phenomenon of interaction surface between antibody and antigen is different, which can be described by different distance ranges between interface residue and antigen. In this paper, given 1 ktr, when the range is chosen as 8, 10 and 12 Å, the amount of interface residues are derived 2, 5 and 10, respectively. Figure 1 graphically depicts the data comparison of the three groups derived from the different distance ranges on 1 ktr. There is clear show the interface residue position in interaction surface relative to antigen in 3D. It will help us to understand the influence of interface residue belong to different rang.

In the experiments, three groups of interface residue data are generated from complex structure. When the distance range is set to 8 Å, we get 329 interface residues. Among them, 253s antigens are protein class and the other 76s antigens are non-protein class. When the distance is set to 10 Å, we have 508 interface residues, where 400s antigens are protein class and the other 108s antigens are non-protein class. When the distance is set to 12 Å, we obtain 668 interface residues with 532s antigens identified as protein class and the rest 136s antigens as non-protein class.

2.3 Feature for inferring distance range

Recent studies in interface residues show that, compared with the physicohemical property of structure, sequence feature can better describe the structural information. This is because the sequence feature has a strong connection of the physicohemical properties of residues. Therefore, sequence feature of target residue is used for identifying distance ranges between interface residue and antigen in this paper.

There are many studies on composing sequence feature of the target residue [28, 29]. One method is surface patch. The target residue is defined as a central surface residue and some nearest surface neighbors are selected. The number of nearest surface neighbors is decided by the surface patch size. Surface patch concerns the neighbor relationship in space. In this paper, sequence patch is used for composing sequence feature. When the sequence patch size is set to be 5, it means the join sequence feature is composed by the target residue and 5 front neighbors and 5 back neighbors (the total of 11 residues) in sequence. Sequence patch is about the neighbor relationship in sequence.

Sequence feature for inferring distance range is coded as follows: each residue is represented by a 20D

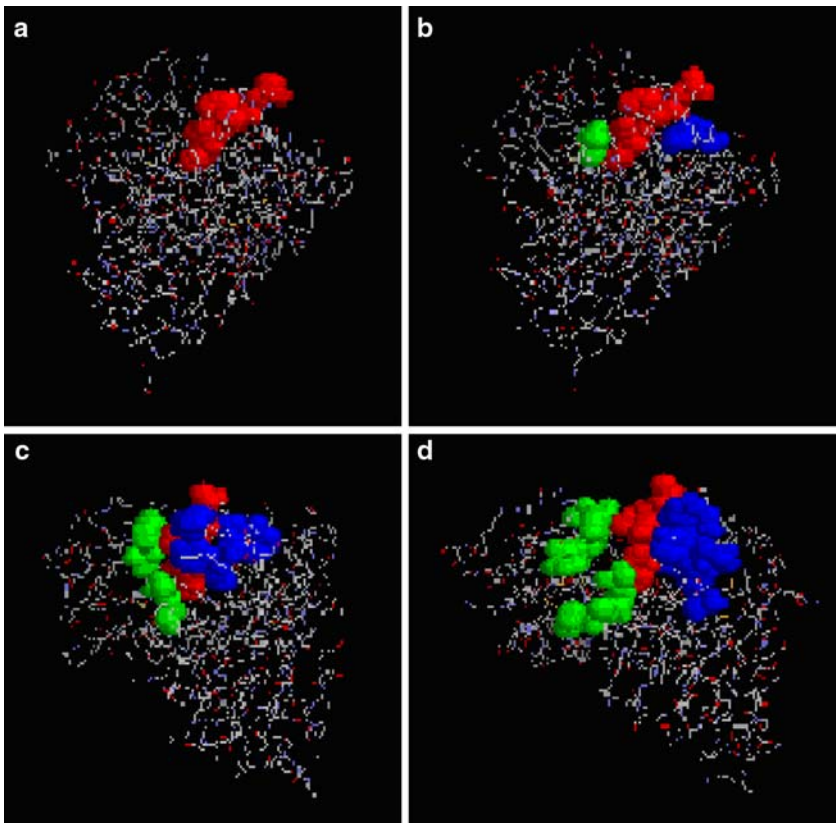

Fig. 1 The figure depicts the interface residues with different distance of complex 1 ktr. **a** Depicts the position of antigen in antibody–antigen complex. From **b** to **d**, the distance range is set to 8, 10, 12 Å, respectively

basic vector, i.e., each kind of residue corresponds to 1 of 20 dimensions in the basic vector. The element of the vector having value one means that it belongs to that kind of residue. Only one position has value one and others has zero. This sequence feature will be coded as a 220D vector if the sequence patch size is set to 5.

2.4 Feature for inferring antigen class

Feature for inference of antigen class is constructed by interface residues belong to different range in antibody structure. It is a 21D vector, which denotes 20 kinds of residue composition in the interface residues plus the number of interface residue belong to distance ranges 8, 10 and 12 Å, respectively.

2.5 Support vector machines

To predict the distance range between antibody's interface residue and antigen from antibody/antigen complex, we adopted algorithm called support vector machine (SVM) [30, 31]. SVM is a kind of learning machine based on statistical learning theory [32, 33]. It is a theory of machine learning focusing on small sample data based on the structural risk minimization principle from computational learning theory [34].

Here is a brief description of the SVM algorithm. Consider the problem of separating the set of training vectors belonging to two separate classes. $D = \{(x^1, y^1),..., (x^l, y^l)\}$, $x \in R^n$, $y \in \{-1,1\}$, with a hyper plane, $(\omega \cdot x) + b = 0$. Figure 2 is a simple linearly separatable case. Solid points and circle points represent two kinds of sample separately. H is the separating line. H_1 and H_2 are the closest lines parallel to the separating line of the two-class sample vectors. The distance between H_1 and H_2 is called margin. The separating hyperplane is said to be optimal if it classifies the samples into two classes without error (training error is zero) and the margin is maximal. The sample vectors in H_1, H_2 are called support vectors. The separating hyperplane equation is $(\omega \cdot x) + b = 0$, where the sample vectors (x_i, y_i), $i = 1,..., n$, should satisfy

$$y_i[(\omega \cdot x) + b] \geq 1, \; i = 1, \ldots, n. \tag{1}$$

The distance of point x to the hyperplane (ω, b) is $d(\omega, b; x) = \frac{|(\omega \cdot x^l) + b|}{\|\omega\|}$. The optimal hyperplane is given by maximizing the margin d, subject to (1). The margin can be given by $\rho(\omega, b) = \frac{2}{\|\omega\|}$. Hence the hyperplane that optimally separates the data is the one that minimizes

$$\phi(\omega) = \frac{\|\omega\|}{2}. \tag{2}$$

The Lagrange function of (2) under constraints (1) is,

$$\phi(\omega, b, a) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l} \alpha_i \left(y^i \left[(\omega, x^i) + b\right] - 1\right), \tag{3}$$

The optimal classification function, if solved, is $f(x) = \text{sgn}((\omega^* x) + b^*)$.

For nonlinear case, we map the original space into a high dimension space by a nonlinear mapping, in which an optimal hyperplane can be sought. The inner product function enables the classification in the new space; however, the computation complexity will not increase. Thus the corresponding program is,

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) * \tag{4}$$

The corresponding separating function is,

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha^*_i y_i K(x_i, x) + b^*\right) \tag{5}$$

This is the so-called SVM.

Support vector machine provides a method to solve the possible dimension disaster in the algorithm: when constructing a discriminant function, SVM does not obtain solution in the feature space after mapping the original sample space into a high dimension space by nonlinear mapping. Instead, it compares the sample vectors in the input space, then it performs nonlinear mapping after the comparison. Function K is called the kernel function of dot product. In [30], it is defined as a distance between sample vectors. The method above can assure all training samples are accurately classified. That is, on condition that the empirical risk is zero, SVM can get the best generalization ability by

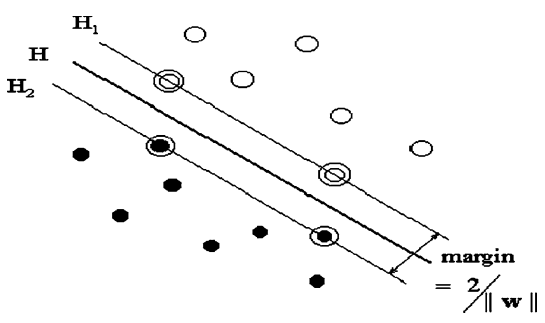

Fig. 2 Optimal separating hyperplane

maximizing the margin. SVM have been used to handle many problems in bioinformatics [10, 18, 35–37].

In this paper, an integrated software for support vector classification named LIBSVM (Version 2.71) [38] is employed to predict antibody interaction sites and antigen class.

2.6 Evaluation measure

The tests of classification accuracy can be divided into two parts: self-consistent test and cross-validation test, which are the common ways in testing the capability of the prediction power. Self-consistent test aims at testing the self-consistence, which takes the same training dataset as testing data. Cross-validation is the test of the generalization ability of the method.

In the self-consistent test, the dataset acts as both the training set and the testing set. The modification process of the parameter in the LIBSVM continues until the result of the self-consistent is satisfied. The indictors, such as prediction accuracy and correlation coefficient, can be obtained from the self-consistent test for analyzing the effectiveness of the method. The rationale behind the method is that using the original data as the testing data, the model shows good prediction accuracy if it is good. Correlation coefficient falls into the range of (–1, 1). The introduction of correlation coefficient is to avoid the negative impacts of the imbalance between different classes of data. For example, if two types of data take up a 4:1 position in a single dataset, then the prediction of the type with a large size of data will be 80% accurate. If the same dataset is used in the testing, then it is meaningless in the case of prediction. When using the model constructed on prediction of the data, the correlation coefficient will be –1 if the prediction is completely contrary to the exact value, 1 if the prediction is correct, and 0 if the prediction is randomly produced. The Correlation coefficient are calculated as follows.

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

where TP (true positive): the number of records in the first class that has been classified correctly.

FP (false positive): the number of records in the second class that has been classified into the first class.

TN (true negative): the number of records in the second class that has been classified correctly.

FN (false negative): the number of records in the first class that has been classified into second class.

After passing the re-substitution test, a fivefold cross validation is applied on the dataset. The details of a fivefold cross validation is discusses as follows. The dataset is split into five parts. One of five acts as the testing set and the other four as the training set to construct the mathematical model. The process rotates for five times with each part as a testing set in a single round. In the generalization test, the mean value of the accuracy in the fivefold cross validation tests is used as the accuracy measure of the experiment. If the utilized method is correct, then the extracted features can well explain the correlation between antibody–antigen interaction residue and antibody–antigen interaction surface. In this situation, the cross-validation test should have a high performance level. The mean accuracy of the fivefold cross validation should also be comparatively high.

3 Results and discussions

3.1 Predict the distance range

3.1.1 Self-consistent test

When the distance range is set to 8 Å and the antigen is of the protein class, the accuracy of the self-consistent test increases from 97.53 to 98.79%; correlation coefficient is from 0.587861 to 0.79216, with the increase of sequence path size from 1 to 5. When sequence path size is four, three indictors: accuracy and correlation coefficient reach their best value, which are 99.01% and 0.829109, respectively.

The self-consistent test results for predicting the distance ranges when the antigen is of protein class, the non-protein class and the class of protein and non-protein can be similarly explained as shown in Tables 1, 2 and 3, respectively.

3.1.2 Cross-validation test

When the distance range is 8 Å and the antigen is of the protein class, the accuracy of the cross validation on the data increases from 94.66 to 95.79%, with the increase of sequence path size from 1 to 5. When sequence path size is four, the accuracy reaches its best value of 95.86%. When the antigen is of the non-protein class, the accuracy of the cross validation increases from 93.72 to 94.52%, with the changes of the sequence path size from 1 to 5. When sequence path size is four or five, the best accuracy value is 94.52%. When the

Table 1 The self-consistent test results with different sequence patch size when the antigen is protein class

Distance range (Å)	Sequence patch size	Accuracy (%)	Correlation coefficient
8	1	97.53	0.587861
8	2	98.96	0.820909
8	3	98.72	0.77995
8	4	99.01	0.829109
8	5	98.79	0.79216
10	1	97.03	0.672668
10	2	98.57	0.839235
10	3	98.96	0.882228
10	4	98.67	0.849841
10	5	98.33	0.812528
12	1	95.43	0.613591
12	2	98.12	0.834975
12	3	97.44	0.777317
12	4	98.09	0.832873
12	5	97.68	0.797929

Table 2 The self-consistent test results with different sequence patch size when the antigen is non-protein class

Distance range (Å)	Sequence patch size	Accuracy (%)	Correlation coefficient
8	1	96.77	0.515651
8	2	99.01	0.846579
8	3	98.65	0.79309
8	4	98.65	0.796243
8	5	99.46	0.918223
10	1	96.41	0.60911
10	2	94.88	0.450205
10	3	95.51	0.521592
10	4	98.56	0.840512
10	5	98.11	0.800671
12	1	97.22	0.751736
12	2	98.65	0.878701
12	3	99.10	0.917999
12	4	98.38	0.853383
12	5	98.56	0.869748

antigen is of the protein and non-protein class, the accuracy of the cross validation increases from 94.34 to 95.83%, with the increase of the sequence path size from 1 to 5. When sequence path size is five, the accuracy reaches its best value of 95.83%.

The accuracy of the cross validation for the distance ranges 10 and 12 Å when the antigen is of the protein class, non-protein class, and the class of protein and non-protein can be similarly explained as shown in Table 4.

3.2 Identification of antigen class

The experiment of identifying the antigen class by composing antibody interface residues belong to different range is designed in two aspects. First, the

Table 3 The self-consistent test results with different sequence patch size when the antigen is protein and non-protein class

Distance range (Å)	Sequence patch size	Accuracy (%)	Correlation coefficient
8	1	97.39	0.573704
8	2	98.88	0.811367
8	3	98.50	0.74865
8	4	98.91	0.817575
8	5	98.69	0.780672
10	1	96.80	0.649831
10	2	98.40	0.820558
10	3	98.08	0.785489
10	4	98.88	0.873085
10	5	98.17	0.796534
12	1	95.51	0.616922
12	2	98.00	0.823716
12	3	97.93	0.817578
12	4	97.51	0.781856
12	5	97.70	0.797733

Table 4 The cross validation test results with different sequence patch size and antigen class

Distance range (Å)	Sequence patch size	Antigen: protein (%)	Antigen: non-protein (%)	Antigen: protein and non-protein (%)
8	1	94.66	93.72	94.34
8	2	95.43	93.72	94.95
8	3	95.65	94.25	95.35
8	4	95.86	94.52	95.60
8	5	95.79	94.52	95.83
10	1	93.14	93.72	93.30
10	2	94.63	93.72	93.83
10	3	94.85	94.25	94.67
10	4	95.50	94.52	95.31
10	5	95.65	94.52	95.41
12	1	91.76	92.01	92.00
12	2	93.57	92.10	93.95
12	3	93.94	93.36	94.46
12	4	94.59	94.08	95.03
12	5	95.24	94.70	95.45

The accuracy in the table denotes the average accuracy of five-fold experiment

interface residues of antibody–antigen interaction belong to different distance range are computed directly from its structure information. Then, we use the SVM to train and test the input features that consist of the residue composition of antibody interface residues and the total of interface residues. The prediction results are listed in Table 5 with different distance ranges and they show that high accuracy is attained when the distance range is set to 8 and 12 Å.

Another test is performed through constructing a two-layer classifier. Using prediction results of interface residue's different distance range as the start point of prediction of antigen class, we can examine the

Table 5 The self-consistent and cross validation test results with different distance threshold

Distance range (Å)	Self-consistent test accuracy (%)	Self-consistent test correlation coefficient	Cross validation test average accuracy of fivefold experiment (%)
8	100	1	86.49
10	94.59	0.669643	83.78
12	100	1	86.49

antibody sequence and surface residues information, and predict antigen class directly. The reason is that the high prediction accuracy of interface residues is the solid guarantee and makes this scheme feasible. As evidence, the prediction accuracy listed in Table 6 is satisfactory.

4 Discussions

The vast quantities of existing immunological data and advanced information technology have boosted the research work on computational immunology. Understanding the circumstantialities of antibody–antigen interaction surface via information technology becomes a new research direction in immunoreaction. Experimental data analysis by using machine learning methods may help explain and provide significant insight into the complex phenomenon of antibody–antigen interaction. The circumstantialities of antibody–antigen interaction surface can be observed by the distance between antibody's interface residue and antigen. It will help us to understand position of each interface residue relative to antigen in 3D, which connect affinity of antibody–antigen interaction.

As a pilot study of discovering the correlation between antibody–antigen interaction residue and antibody–antigen interaction surface, total of 668 interface residues was extracted from these 5,253 surface residues and divided into three classes by different distance range. Based on different distance range, sequence patch size and antigen class, 45 samples have been created in the research of prediction distance range between antibody's interface residue and antigen by feature of antibody's surface residue sequence and 3 samples have been chosen for identifying antigen class from the antibody's interface residue character by different distance range. The result shows that different settings have complicate results for finding the prediction accuracy.

Table 6 The cross validation test results with different distance range and sequence patch size combining the antibody interface residue prediction results

Sequence patch size	Distance range (%)		
	8 Å	10 Å	12 Å
1	81.59	78.17	79.57
2	82.12	78.61	81.26
3	82.47	79.31	81.70
4	82.68	79.85	82.19
5	82.88	79.93	82.55

The accuracy list in the table is the average accuracy for fivefold experiment

Through these experiments by using support vector machine, we can observe the following: antigen class greatly influences the accuracy of prediction interface residue's distance range. High accuracy will be achieved by antigen classification. In the same distance range, the increase in the sequence path size will help increase the prediction accuracy. The larger the sequence patch size is, the closer the effect of different distance range on the accuracy will be. When the sequence patch size is small, different distance ranges have a great influence of the accuracy. The results of this paper show that it is possible to infer distance range between antibody's interface residue and antigen and compatible antigen class from the antibody structures data.

The results obtained in this paper can be further explored through the following aspects: The size of manually selected data is still smaller and the balance dataset structure is needed. The prediction accuracy could be lower than what obtained in this paper if the data size is larger with a balanced structure. Interface of antibody combine antigen is difference for different antigen class. We will focus our study on one kind of antigen. The position of every interface residue relative to antigen in 3D is mate affinity of antibody–antigen interaction. More precision of identifying interface residue's distance ranges may provide a chance to derive a batter understanding of interaction between antibody and antigen. If other known computational methods, such as neural networks is used, we may investigate which method bring better results in relationship between antibody structure and antibody functions.

Note that only the sequence information of antibody has been used in this paper. Following this paper, our

future interest will be using different machine learning techniques, such as multiple criteria mathematical programming [39, 40], on the large-scale structure information for possible higher degree of prediction accuracy.

5 Conclusions

For high affinity of antibody–antigen interaction, it is important to know whether two surfaces match each other. The circumstantialities of antibody–antigen interaction are represented by the distance between antibody's interface residue and antigen. In the past, much effort has focused on characters of antibody structure, antibody–antigen binding site, the affinity and specificity of the antibody. This paper has proposed a machine learning approach to explore the interaction between antibody's interface residue and antigen. That is to say, the interaction information can be predicted from the structure characters. Based on the prediction results of different distance ranges, another experiment is performed to predict the class of antigen. Two involved factors, the sequence patch size and the antigen classes, have been studied to show how they affect the prediction accuracy. It has described that the prediction accuracy is dependent on these factors. The findings indicate that there are positive correlations between sequence information of antibody and their effect of combine antigen. Taken together, the results of our research have been shown the machine learning algorithm can be a useful tool for research on the relationship between antibody function and structure from PDB. We shall continue to promote this approach so that it can be adopted by other bioinformatics filed.

Acknowledgments This research has been partially supported by a 973 Project grant (2004CB720103) from the Ministry of Science and Technology, China and the grants (70531040, 70472074) from the National Natural Science Foundation, China. We would like to express our thanks to Mrs. Li Zhang, Northeastern University, USA and Mr. Gang Kou, University of Nebraska, USA for their constructive comments.

References

1. Petrovsky N, Brusic V (2002) Computational immunology: the coming of age. Immunol Cell Biol 80:248–254
2. Webster DM, Henry AH, Rees AR (1994) Antibody–antigen interactions. Curr Opin Struct Biol 4:123–129
3. Stanfield RL, Fieser TM, Lerner RA, Wilson IA (1990) Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 Å. Science 248:712–719
4. Bath TN, Bentley GA, Fischmann TO, Boulot G, Poljak RJ (1990) Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. Nature 347:483–485
5. Colman PM, Laver WG, Varghese JN, Baker AT, Tulloch PA, Air GM, Webster RG (1987) Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. Nature 326:358–363
6. Xiang J, Sha Y, Prasad L, Delbaere LTJ (1996) Complementarity determining region residues aspartic acid at H55 serine at tyrosines at H97 andL96 play important roles in the B72.3 antibody-TAG72 antigen interaction. Protein Eng 9:539–543
7. Chothia C, Lesk AM, Gherardi E, Tomlinson IM, Walter G, Marks JD, Lewelyn MB, Winter G (1992) Structural repertoire of the human Vh segments. J Mol Biol 227:799–817
8. Iba Y, Hayshi N, Sawada I, Titani K, Kurosawa Y (1998) Changes in the specificity of antibodies against steroid antigens by introduction of mutations into complementarity-determining regions of Vh domain. Protein Eng 11:361–370
9. Rees AR, Staunton D, Webster DM (1994) Antibody design: beyond the natural limits. Trends Biotechnol 12:199–207
10. Minakuchi1 Y, Konagaya A (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 17:165–173
11. Chakrabarti P, Janin J (2002) Dissecting protein–protein recognition sites. Proteins 47:334–343
12. Glaser F, Steinberg DM, Vakser A, Ben-Tal N (2001) Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 43:89–102
13. Lu L, Lu H, Skolnick J (2003) Development of United Statistical Potentials describing protein–protein interactions. Biophys J 84:1895–1901
14. Fariselli P, Casadio R (1999) Neural network based predictor of residue contacts in proteins. Protein Eng 12:15–21
15. Fariselli P, Pazos F, Valencia A, Casadio R (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269:1356–1361
16. Ofran Y, Rost B (2003) Analysing six types of protein–protein interfaces. J Mol Biol 325:377–387
17. Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544:236–239
18. Yan C, Honavar V, Dobbs D (2004) Identification of interface residues in protease-inhibitor and antigen–antibody complexes: a support vector machine approach. Neural Comput Appl 13:123–129
19. Yan C, Dobbs D, Honavar V (2004) A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 20(Suppl 1):i371–i378
20. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226
21. Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. Protein Eng 3:659–665
22. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA (2001) Prediction of protein surface accessibility with information theory. Proteins 42:452–459
23. Li X, Pan XM (2001) New method for accurate prediction of solvent accessibility from protein sequence. Proteins 42:1–5
24. Pascarella S, De Persio R, Bossa F, Argos P (1998) Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins 32:190–199
25. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
27. Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20
28. Jones S, Thornton JM (1997a) Analysis of protein–protein interaction sites using surface patches. J Mol Biol 272:121–132
29. Jones S, Thornton JM (1997b) Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 272:133–143
30. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge
31. Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–167
32. Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20:273–297
33. Bradley PS, Fayyad UM, Magasarian OL (1999) Mathematical programming for data mining: formulations and challenges. INFORMS J Comput 11:217–238
34. Li J, Liu J, Xu W, Shi Y (2004) Support vector machines approach to credit assessment. In: Sloot PMA et al (eds) ICCS 2004, LNCS 2658, Springer, Berlin Heidelberg New York, pp 892–899
35. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. Proc Natl Acad Sci 97:262–267
36. Furey T, Cristianini N, Duffy N (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914
37. Haussler D (1999) Convolution kernels on discrete structures, Tech Rep UCSC-CRL-99–10, UC Santa Cruz
38. http://www.csie.ntu.edu.tw/~cjlin/libhttp://www.csie.ntu.edu.tw/~cjlin/libsvm
39. Kou G, Liu X, Peng Y, Shi Y, Wise W, Xu W (2003) Multiple criteria linear programming to data mining: models, algorithm designs and software developments. Optim Methods Softw 18:453–473
40. Zheng J, Zhuang W, Yan N, Kou G, Peng H, et al (2004) Classification of HIV-1 mediated neuronal dendritic and synaptic damage using multiple criteria linear programming. Neuroinformatics 2:303–326