國立交通大學

生物資訊研究所

碩士論文

從蛋白質序列預測殘基相對溶劑可接觸性

Prediction of Protein Relative Solvent Accessibility from Amino Acid Sequence

研究生：徐蔚倫

指導教授：黃鎮剛　教授

中華民國九十四年七月

從蛋白質序列預測殘基相對溶劑可接觸性

Prediction of Protein Relative Solvent Accessibility from Amino Acid Sequence

研究生：徐蔚倫　　Student：Wei-Lun Hsu

指導教授：黃鎮剛　　Advisor：Jenn-Kang Hwang

國立交通大學
生物資訊研究所
碩士論文

A Thesis
Submitted to Institute of Bioinformatics
College of Biological Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
In

Bioinformatics

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

從蛋白質序列預測殘基相對溶劑可接觸性

學生：徐蔚倫　　　　　　　　　　　　　　　指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

中文摘要

從序列資訊來預測蛋白質三級結構是目前生物學研究上非常重要的目標之一，而正確的預測蛋白質相對溶劑可接觸性則可以提供蛋白質三級結構相關的資訊。蛋白質相對溶劑可接觸性(RSA)代表著蛋白質上某一個氨基酸和溶劑接觸的程度。通常蛋白質的結合處會位於它的表面，因此，若能正確的預測蛋白質位於表面的氨基酸位置，就能夠更進一步的瞭解該蛋白質的功能。此外，一個蛋白質位於表面和包埋在蛋白質內部的氨基酸分佈，也被觀察到和蛋白質在細胞內的位置有很大的關連性。

我的方法是利用支持向量機將局部和整體的蛋白質資訊，其中最好的結果是利用位置加權矩陣(PSSM)、二級結構特徵值(secondary structure profile)和氨基酸親水程度(hydropathy indexes)作為輸入向量。這個方法對於RS126 資料群在以 25%為分類閾值時，可以達到 77.2%，和最近幾年的在這方面的研究成果 75%-78.3%達到相近的程度，而在將 RSA 分成十類的預測結果中也可達到 15.2% 平均絕對誤差，預測值和實驗值達到 0.51 的相關性。

Prediction of Protein Relative Solvent Accessibility from Amino Acid Sequence

Student：Wei-Lun Hsu　　　　Advisor：Jenn-Kang Hwang

Institute of Bioinformatics, National Chiao Tung University

Abstract

The prediction of the three-dimensional structure from its sequence is probably one of the most important goals of modern biology. The accurate prediction of protein relative solvent accessibility is useful for the prediction of tertiary structure of a protein. Amino acid solvent accessibility is the degree to which a residue in a protein is accessible to a solvent molecule. Because the binding sites of a protein are usually located on its surface, accurately predicting the surface residues can be regarded as an important step toward determining its function. On the other hand, it has been observed that the distribution of surface residues of a protein is correlated with its subcellular environments; consequently, information of surface residues may improve the prediction of protein subcellular localization.

Presently, out best method is based on the support vector machines using as the input feature vectors, the sliding window that includes the local environment descriptors such as PSSM, secondary structure profile and hydropathy indexes. In my work, relative solvent accessibility based on a 2-state model, for 25%, 16%, 5%, and 0% accessibility are predicted at 77.2%, 77.1%, 80.4%, and 88.4% accuracy, respectively. Furthermore, solvent accessibility prediction methods have in recent years reached accuracy in the range of 75.0-78.3% at 25% threshold. And the results based in a 10-state model can reach 15.2% mean absolute error and 0.51 correlations.

誌 謝

在求知的路上總是要一直不斷的經歷各種學習的過程，才能不斷的成長進步。在研究所二年的時間裡，各方面我都是收穫良多，真的很慶幸能身處在這樣一個優質的學習環境中體驗研究生生活。

首先，很感謝我的指導教授黃鎮剛老師，帶領我進入研究的領域，不但提供我們一個良好獨立的研究環境，更不斷的在我研究遇到問題、瓶頸時，給予我適當的建議與協助。在老師身上，我學習到的不只是研究方面的知識和方法，更重要的是一個學者對求知的熱誠和執著態度。

另外，也很謝謝實驗室的夥伴們，除了生活上的彼此協助照應、精神上的鼓勵之外，很多研究上的問題也是透過大家的討論和互相協助，才能夠順利的解決，讓我的研究能如期完成。

最後，我也要感謝我的家人，在我求學路上，一直全力支持我，讓我能夠在研究上全力以赴、順利完成學業。僅以此論文作為我碩士研究生涯的總結，並將此獻給所有關心我及我所誠摯感謝的人。

目錄

圖目錄

表 目 錄

1. Introduction

Knowledge of protein’s three-dimensional (3D) structure is essential for a full understanding of its functionality. However, only a small fraction of the enormous number of sequenced proteins has their structure determined. In order to reduce the gap between sequence and structure, developing reliable and applicable structure prediction methods has become a more important task in computational biology. Thus simplification of the problem from 3D structures to 1D feature may be useful as a first-step. The relative solvent accessibility of an amino acid in protein is a real value that represents the solvent exposed surface area of this residue in relative terms. The prediction of secondary structure is more familiar and well-defined aspect of the problem and the prediction of residue solvent accessibility is another aspect. Secondary structures and solvent accessibilities of amino acid residues give a useful insight into the structure and function of a protein. In particular, the knowledge of solvent accessibility has assisted alignments in regions of remote sequence identity for threading.

In addition to providing insight into the conformation of 3D structure, prediction of residue solvent accessibility has many other applications. For example, it has been observed that the distribution of surface residues of a protein is correlated with its subcellular environment. Many studies[1] have suggested that hydrophobic core residues are likely sites of deleterious mutations. The residues in site of deleterious mutations may be critical for protein stability. Wang and Molt[2] found that the vast majority of disease mutations affected protein stability rather than function and could be predicted using straightforward rules.

There are various approaches have been developed, typically by examining a window of residues centered at the test residue and using sequence input or sequence profile as input vectors. These include neural network[3-7], Bayesian statistics[8], linear regressions[9, 10], information theory[11], support vector machines[12, 13] and fuzzy k-nearest neighbor method[14]. However, in contract to the secondary structure, there is no widely accepted criterion for classifying the experimentally determined solvent accessibility into a finite number of discrete states such as buried, intermediate and exposed states. Also, the prediction accuracies of solvent accessibility are lower than those for secondary structure prediction, since the solvent accessibility is less conserved than secondary structure[3], although there has been some progress recently.

In this work, relative solvent accessibility based on a 2-state model in RS126 dataset, for 25%, 16%, 5%, and 0% accessibility are predicted at 77.2%, 77.1%, 80.4%, and 88.4% accuracy, respectively. And the results based in a 10-state model can reach 15.2% mean absolute error and 0.51 correlations.

2. Material and Methods

2.1 Definition and thresholds of solvent accessibility

The relative solvent accessibility of an amino acid is the degree that present the residue in a protein is accessible to solvent molecules. The relative solvent accessibility can be calculated by the formula as follows,

$$\mathrm{Re}\,lACC = 100 \cdot \frac{ACC}{MaxACC} \tag{1}$$

RelACC : relative solvent accessibility

ACC : solvent accessible surface areas

MaxACC : maximum value of solvent accessible surface areas of each kind of residue

Where ACC of ith residue is the solvent accessible surface areas calculated by the dictionary of protein secondary structure program (DSSP) [15], and the MaxACC are taken from MOLSIM program (Table 1). RelACC is derived from normalizing ACC by dividing the DSSP accessibility by the maximum accessibility of amino acid residues corresponding to accessibility for a Gly-X-Gly extended tripeptide conformation. RelACC can hence adopt values between 0% and 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue, respectively.

Various thresholds have been used to classify residues as buried and exposed (2-state prediction) or buried, intermediate and exposed (3-state prediction) in previously published results. In this work, thresholds of 0, 5, 16, 25% for buried/exposed in the 2-state predictions, thresholds of 9, 36% for buried/intermediate/exposed in the 3-state predictions and 10, 20, 30, 40, 50, 60, 70, 80, 90% for classifying residues in 10 classes are used.

2.2 Training and data sets

The solvent accessibility was predicted for three sets of proteins in order to evaluate the performance of our method. They are a 126-protein dataset of Rost and Sander[3] (Table 2), a set of 215 protein chains selected by Naderi-Manesh[11] (Table 3), and 480 proteins from Kim and Park[13] (Table 4). Each of these data sets consists of protein sequence with less than 25% homology. To perform 7-fold cross validation test, we randomly divided these data sets into 7 groups, each containing a roughly equal number of protein sequences. One group was chosen as the testing set, while others were merged into the training set.

2.3 Feature vectors

In the first phase, we have individually benchmarked 8 different local descriptor types in prediction experiment in RS126 dataset. Our results revealed that the five best-performing sets of local descriptors were using PSSM (position-specific scoring matrices), secondary structure profiles, hydropathy indexes, amino acid composition and amino acid properties by their R groups. In second phase, the five top-performing local descriptors were combined to obtain a new method with improved performance. And then, we add 2 different global descriptors types into the method to see if global descriptors are useful for solvent accessibility prediction.

2.3.1 Local descriptors

We represent the local environment of sequences by using sliding window coding scheme. Increasing the window size can provide more local information and it is reasonable to expect that prediction accuracy would increase with the enlargement of the window size. But the past research[12] found that window size has a very limited effect on prediction accuracy.

Therefore, we followed the paper and selected 15 for the window size for the prediction of protein relative solvent accessibility

2.3.1.1 PSSM (position-specific scoring matrices)

We use PSI-BLAST[16] against non-redundant protein sequence database for 5 iterations to produce PSSM, which has 20*m elements, where m is the length of query sequences and each element represents the log-likelihood of that specific residue substitution at that position in the template (based on a weighted average of BLOSUM62 matrix scores for the given alignment position). PSI-BLAST uses a simple but effective scheme for weighting the contribution of locally different numbers of sequences to the resulting profiles. The profile matrix elements are scaled to the required 0-1 range by using the standard logistic function as below, where x of the equation is the raw profile matrix value.

$$Score = \frac{1}{1+e^{-x}} \qquad (2)$$

Then, the input vector has 20*15 components (Illustrated in Figure 1). Each component can be expressed as $PSSM_{ijk}$, where i is the predicted residue, j is the location in the sliding window and k is the type of amino acid.

2.3.1.2 Secondary structure profiles

The secondary structure profile has 3*m elements. The elements describe the probabilities of each residue in the three kinds of secondary structures which are Helix, Sheet and Coil. These profiles were generated from the PSIPRED[17] program, and each input vector has 3*15 components (Illustrated in Figure 2). Each component can be expressed as SS_{ijk}, where i is the predicted residue, j is the location in the sliding window and k is the type of secondary structure including helix, strand and coil.

2.3.1.3 Hydropathy indexes

In this study, we use the hydropathy indexes to represent each residue (Table 5). In general, the hydropathy indexes can be used to measure the tendency of an amino acid to seek an aqueous environment or a hydrophobic environment. The hydropathy indexes refer to the book “Lehninger Principles of Biochemistry”. Each input vector has 1*15 components (Illustrated in Figure 3). Each component can be expressed as H_{ij}, where i is the predicted residue and j is the location in the sliding window.

2.3.1.4 Amino acid in binary coding

We code the protein sequences in binary mode, such as “1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0” represents A (alanine). Each input vector has 20*15 components (Illustrated in Figure 4). Each component can be expressed as AA_{ijk}, where i is the predicted residue, j is the location in the sliding window and k is the type of amino acid.

2.3.1.5 Amino acid properties in binary coding

Knowledge of chemical properties of the standard amino acids is central to an understanding of biochemistry. The chemical properties can be simplified by grouping the standard amino acids into 5 main classes based on the properties of their R groups, and others into the sixth class.(Table 6) The classification refer to the book “Lehninger Principles of Biochemistry”. Each input vector has 6*15 components (Illustrated in Figure 5). Each component can be expressed as $PROP_{ijk}$, where i is the predicted residue, j is the location in the sliding window and k is the property of amino acid.

2.3.1.6 Local amino acid composition

The amino acid percentage is the percentage of each residue in the sliding window. Each input vector has 20 components (Illustrated in Figure 6). Each component can be expressed as LAA_{ik}, where i is the predicted residue and k is the type of amino acid.

2.3.1.7 Local amino acid properties composition

The amino acid properties percentage is the percentage of each residue property in the sliding window. Each input vector has 6 components (Illustrated in Figure 7). Each component can be expressed as $LPROP_{ik}$, where i is the predicted residue, j is the location in the sliding window and k is the property of amino acid.

2.3.1.8 Residue size

We use the volume of amino acid as the input vector (Table 1). Each input vector has 1*15 components (Illustrated in Figure 8). Each component can be expressed as R_{ij}, where i is the predicted residue and j is the location in the sliding window.

2.3.2 Global descriptors

We also represent the global environment of sequences by using global sequence composition and global character composition method.

2.3.2.1 Global amino acid composition

The amino acid percentage is the percentage of each residue in a given protein sequence totally. Each input vector has 20 components (Illustrated in Figure 9). Each component can be expressed as GAA_k, where k is the type of amino acid.

2.3.2.2 Global amino acid properties composition

The amino acid properties percentage is the percentage of each residue property in a given protein sequence totally. Each input vector has 6 components (Illustrated in Figure 10). Each component can be expressed as $GPROP_k$, where k is the property of amino acid.

2.4 Performance measures

In order to compare with other works, we use different measures to evaluate 2-state, 3-state and 10-state prediction methods as below.

2.4.1 In 2-state prediction and 3-state prediction

In this work, two measures are used to evaluate the performance of prediction methods. One is accuracy, the percentage of correctly classified residues, and the other is Matthew's correlation coefficients (MCC). These measures can be calculated by the following equation:

$$Accuracy = \frac{\sum_i^c p_i}{N} \tag{3}$$

$$MCC_i = \frac{p_i n_i - o_i u_i}{\sqrt{(p_i + o_i)(p_i + u_i)(n_i + o_i)(n_i + u_i)}} \tag{4}$$

Where N is the total number of residues, and c is the class number. And $p_{i,}$ n_i, o_i and u_i are the number of true positives, true negatives, false positives and false negatives for class i, respectively. The MCCs have the same value for the two classes in the case of the 2-state prediction.

2.4.2 In 10-state prediction

Here, we applied several different measures, including the mean absolute error (MAE) of prediction defined as the absolute difference between the predicted and experimental value of relative solvent accessibility, per residue.

$$MAE = \frac{\sum \left| (ASA)_{pred} - (ASA)_{\exp} \right|}{N} \tag{5}$$

Where summation is carried out for all residues and N is the total number of predictions. In addition, Pearson's "r" is also used in some places, and it is calculated as the ratio of the covariance between the predicted and experimental relative solvent accessibility as below.

$$r = \frac{N\left(\sum (ASA)_{pred} (ASA)_{\exp}\right) - \left(\sum (ASA)_{pred}\right)\left(\sum (ASA)_{\exp}\right)}{\sqrt{\left[N\sum (ASA)_{pred}{}^{2} - \left(\sum (ASA)_{pred}\right)^{2}\right]\left[N\sum (ASA)_{\exp}{}^{2} - \left(\sum (ASA)_{\exp}\right)^{2}\right]}} \tag{6}$$

2.5 Support vector machine

Support vector machine (SVM), first proposed by Vapnik and co-workers[18] based on statistical learning theory, has quickly become one of the most popular classification and regression methods, due to its flexibility in choosing a similarity function, the ability to handle large feature spaces and accuracy. It has been used in various area, such as protein structure prediction[19], protein fold recognition[20] and microarray data analysis[21].

An SVM finds a nonlinear decision function in the input space by implicitly mapping the data into a linear separable higher dimensional feature space and separating the data there by maximizing the geometric margin and minimizing the training error at the same time. Yuan et al.[5] implemented SVM^{light} to predict solvent accessibility. Here, we use LIBSVM version 2.6, which is a multi-class SVM developed by Chang and Lin.[22]

3. Server Development and Administration System Overview

Figure 2 shows the flowchart of RSA prediction server, and most of the process was built using the Perl language and CGI package. And the prediction result can automatically show on the webpage when the relative solvent accessibility prediction was finished.

4. World Wide Web Interface

4.1 Models trained by the three benchmark data sets

User can select an appropriate training dataset of the three benchmark datasets. The consuming time will increase depending on the dataset size.

4.2 Models of different state thresholds

User can choose a 2-state model (25%), 3-state model (9%;36%) or 10-state model (10%;20%;30%;40%;50%;60%;70%;80%;90%) for different purposes.

4.3 Query sequence

User can input their query sequence in FASTA format or directly upload the FASTA format file to the SAS prediction server.

4.4 Website

The RSA prediction server is available at http://e100.life.nctu.edu.tw/~weilun/. Figure3 shows the World Wide Web interface of RSA prediction server. Figure 4 shows the standard output webpage of prediction.

5. Results and Discussion

5.1 Prediction accuracies and MCC values using individual local descriptors with RS126 dataset in 2-state prediction model

In the first phase, the performance was evaluated in terms of Q_{total}, Q_b, Q_e and MCC (Table 7). Of the 8 local descriptors, 5 were ranging 65.73% to 75.33%. They are PSSM, amino acid composition, secondary structure profiles, character information and hydropathy indexes in order.

5.2 Prediction accuracies and MCC values using comprehensive local descriptors with RS126 dataset in 2-state prediction model

In the second phase, we combined the top 5 best-performing local descriptors and 2 global descriptors in many kinds of combinations. In Table 8 we show the results.

Here, we found that PSSM involved more useful information in relative solvent accessibility prediction. The hydropathy indexes performed worse than secondary structure profiles in the individual test, but when combining with PSSM, the hydropathy indexes could increase the prediction accuracy more than the secondary structure profiles did. The combination of these three input vectors can reach the accuracy rate at 77.23% in RS-126 dataset. However, amino acid composition and amino acid properties also seem to offer duplicated information with PSSM, but the amino acid composition could slightly improve the prediction accuracy about 0.5%. The other feature vectors seem not to give useful information about prediction including global descriptors.

Overall, the best combinations of input vector were PSSM, secondary structure profiles, hydropathy indexes and amino acid composition in this stage. Compared with the individual performance of the 8 different descriptors, the increased accuracy was in the range 68.40% to 77.78%.

5.3 Prediction accuracies and MCC values of the benchmark test on the RS126, NM215 and KP480 datasets in 2-state prediction model and its comparison with other methods

The performances of prediction in this study are showed in Table9 and Table10. In this stage, the result of top-4 outperforming feature vectors is almost the same as the result of top-3 outperforming feature vectors. To reduce the training space, we took the top-3 outperforming feature vectors as our best work. They are PSSM, secondary structure profiles and hydroparthy indexes. Our best work shows 60.2% accuracy for 3-state prediction (9%;36% thresholds) and 88.4%,81.4%,77.9% and 77.2% for the 2-state prediction with thresholds of 0%,5%,16% and 25% on the RS126 dataset, respectively. It shows slightly better prediction than other methods on the RS126 dataset expect for "Fuzzy k-NN".

5.4 MAE and Pearson's "r" values of the benchmark test on the RS126 and KP480 datasets in 10-state prediction model

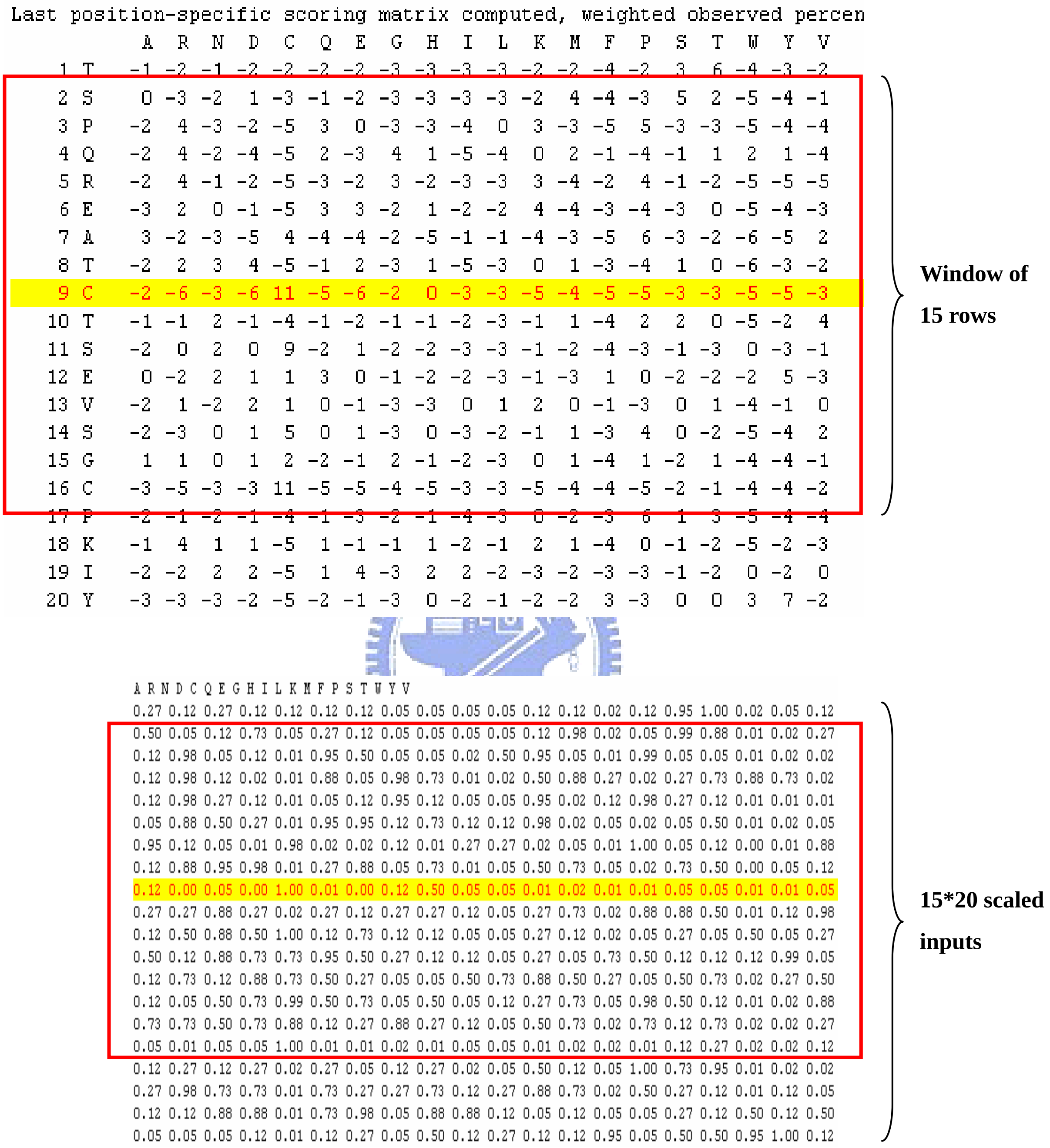
We used the 10-state prediction model to simulate the real value prediction of solvent accessibility prediction. We took the same measured way to compare with the other real value prediction methods. All results were shown in Table11 and our best work achieved comparable results. The MAE value is 15.2 and Pearson's r is 0.506 in RS-126 dataset. We also transform the 10-state prediction into 2-state to give a rough evaluation. At the threshold of 20%, the accuracy achieved 72.8%, 74.4% in RS-126 and KP-480 dataset.

6. Conclusion

In this study, our best method achieved the similar performance as the researches did in the recent years. Hence, we can conclude that the three input vectors in our best work can catch enough information to predict relative solvent accessibility accurately.

Our research shows that local descriptors are sufficient for predicting accurately without global descriptors. Secondary structure profiles and hydropathy indexes both can be complementary to PSSM, but in individual case, hydropathy indexes can only achieve poor performance. Sequence information and character information offer overlapped information with PSSM.

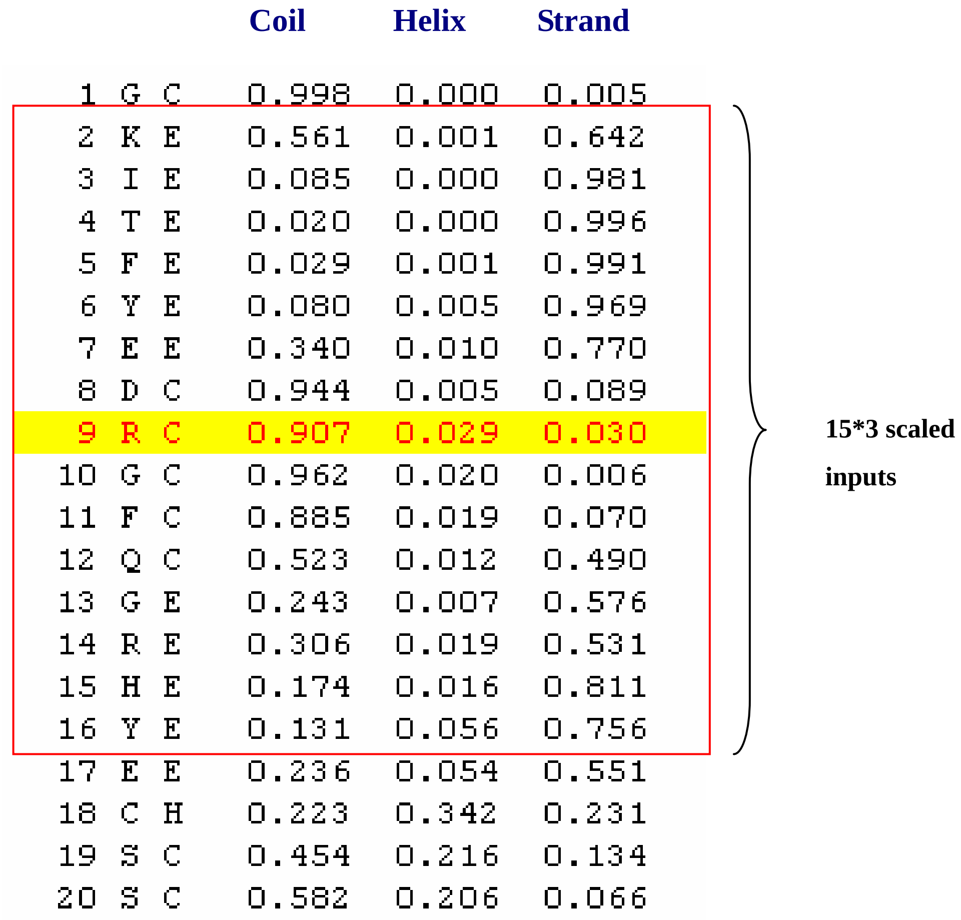
In conclusion, our work achieves good prediction accuracy on the three benchmark datasets in 2-state, 3-state and 10-state model. And in the future work, we can apply our method in real value prediction.



```
Last position-specific scoring matrix computed, weighted observed percen
           A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
   1 T    -1 -2 -1 -2 -2 -2 -2 -3 -3 -3 -3 -2 -2 -4 -2  3  6 -4 -3 -2
   2 S     0 -3 -2  1 -3 -1 -2 -3 -3 -3 -3 -2  4 -4 -3  5  2 -5 -4 -1
   3 P    -2  4 -3 -2 -5  3  0 -3 -3 -4  0  3 -3 -5  5 -3 -3 -5 -4 -4
   4 Q    -2  4 -2 -4 -5  2 -3  4  1 -5 -4  0  2 -1 -4 -1  1  2  1 -4
   5 R    -2  4 -1 -2 -5 -3 -2  3 -2 -3 -3  3 -4 -2  4 -1 -2 -5 -5 -5
   6 E    -3  2  0 -1 -5  3  3 -2  1 -2 -2  4 -4 -3 -4 -3  0 -5 -4 -3
   7 A     3 -2 -3 -5  4 -4 -4 -2 -5 -1 -1 -4 -3 -5  6 -3 -2 -6 -5  2
   8 T    -2  2  3  4 -5 -1  2 -3  1 -5 -3  0  1 -3 -4  1  0 -6 -3 -2
   9 C    -2 -6 -3 -6 11 -5 -6 -2  0 -3 -3 -5 -4 -5 -5 -3 -3 -5 -5 -3
  10 T    -1 -1  2 -1 -4 -1 -2 -1 -1 -2 -3 -1  1 -4  2  2  0 -5 -2  4
  11 S    -2  0  2  0  9 -2  1 -2 -2 -3 -3 -1 -2 -4 -3 -1 -3  0 -3 -1
  12 E     0 -2  2  1  1  3  0 -1 -2 -2 -3 -1 -3  1  0 -2 -2 -2  5 -3
  13 V    -2  1 -2  2  1  0 -1 -3 -3  0  1  2  0 -1 -3  0  1 -4 -1  0
  14 S    -2 -3  0  1  5  0  1 -3  0 -3 -2 -1  1 -3  4  0 -2 -5 -4  2
  15 G     1  1  0  1  2 -2 -1  2 -1 -2 -3  0  1 -4  1 -2  1 -4 -4 -1
  16 C    -3 -5 -3 -3 11 -5 -5 -4 -5 -3 -3 -5 -4 -4 -5 -2 -1 -4 -4 -2
  17 P    -2 -1 -2 -1 -4 -1 -3 -2 -1 -4 -3  0 -2 -3  6  1  3 -5 -4 -4
  18 K    -1  4  1  1 -5  1 -1 -1  1 -2 -1  2  1 -4  0 -1 -2 -5 -2 -3
  19 I    -2 -2  2  2 -5  1  4 -3  2  2 -2 -3 -2 -3 -3 -1 -2  0 -2  0
  20 Y    -3 -3 -3 -2 -5 -2 -1 -3  0 -2 -1 -2 -2  3 -3  0  0  3  7 -2
```

```
A R N D C Q E G H I L K M F P S T W Y V
0.27 0.12 0.27 0.12 0.12 0.12 0.12 0.05 0.05 0.05 0.05 0.12 0.12 0.02 0.12 0.95 1.00 0.02 0.05 0.12
0.50 0.05 0.12 0.73 0.05 0.27 0.12 0.05 0.05 0.05 0.05 0.12 0.98 0.02 0.05 0.99 0.88 0.01 0.02 0.27
0.12 0.98 0.05 0.12 0.01 0.95 0.50 0.05 0.05 0.02 0.50 0.95 0.05 0.01 0.99 0.05 0.05 0.01 0.02 0.02
0.12 0.98 0.12 0.02 0.01 0.88 0.05 0.98 0.73 0.01 0.02 0.50 0.88 0.27 0.02 0.27 0.73 0.88 0.73 0.02
0.12 0.98 0.27 0.12 0.01 0.05 0.12 0.95 0.12 0.05 0.05 0.95 0.02 0.12 0.98 0.27 0.12 0.01 0.01 0.01
0.05 0.88 0.50 0.27 0.01 0.95 0.95 0.12 0.73 0.12 0.12 0.98 0.02 0.05 0.02 0.05 0.50 0.01 0.02 0.05
0.95 0.12 0.05 0.01 0.98 0.02 0.02 0.12 0.01 0.27 0.27 0.02 0.05 0.01 1.00 0.05 0.12 0.00 0.01 0.88
0.12 0.88 0.95 0.98 0.01 0.27 0.88 0.05 0.73 0.01 0.05 0.50 0.73 0.05 0.02 0.73 0.50 0.00 0.05 0.12
0.12 0.00 0.05 0.00 1.00 0.01 0.00 0.12 0.50 0.05 0.05 0.01 0.02 0.01 0.01 0.05 0.05 0.01 0.01 0.05
0.27 0.27 0.88 0.27 0.02 0.27 0.12 0.27 0.27 0.12 0.05 0.27 0.73 0.02 0.88 0.88 0.50 0.01 0.12 0.98
0.12 0.50 0.88 0.50 1.00 0.12 0.73 0.12 0.12 0.05 0.05 0.27 0.12 0.02 0.05 0.27 0.05 0.50 0.05 0.27
0.50 0.12 0.88 0.73 0.73 0.95 0.50 0.27 0.12 0.12 0.05 0.27 0.05 0.73 0.50 0.12 0.12 0.12 0.99 0.05
0.12 0.73 0.12 0.88 0.73 0.50 0.27 0.05 0.05 0.50 0.73 0.88 0.50 0.27 0.05 0.50 0.73 0.02 0.27 0.50
0.12 0.05 0.50 0.73 0.99 0.50 0.73 0.05 0.50 0.05 0.12 0.27 0.73 0.05 0.98 0.50 0.12 0.01 0.02 0.88
0.73 0.73 0.50 0.73 0.88 0.12 0.27 0.88 0.27 0.12 0.05 0.50 0.73 0.02 0.73 0.12 0.73 0.02 0.02 0.27
0.05 0.01 0.05 0.05 1.00 0.01 0.01 0.02 0.01 0.05 0.05 0.01 0.02 0.02 0.01 0.12 0.27 0.02 0.02 0.12
0.12 0.27 0.12 0.27 0.02 0.27 0.05 0.12 0.27 0.02 0.05 0.50 0.12 0.05 1.00 0.73 0.95 0.01 0.02 0.02
0.27 0.98 0.73 0.73 0.01 0.73 0.27 0.27 0.73 0.12 0.27 0.88 0.73 0.02 0.50 0.27 0.12 0.01 0.12 0.05
0.12 0.12 0.88 0.88 0.01 0.73 0.98 0.05 0.88 0.88 0.12 0.05 0.12 0.05 0.05 0.27 0.12 0.50 0.12 0.50
0.05 0.05 0.05 0.12 0.01 0.12 0.27 0.05 0.50 0.12 0.27 0.12 0.12 0.95 0.05 0.50 0.50 0.95 1.00 0.12
```

Figure 1. An outline of the PSSM input vectors, which shows how the PSI-BLAST score matrices are processed.



			Coil	Helix	Strand
1	G	C	0.998	0.000	0.005
2	K	E	0.561	0.001	0.642
3	I	E	0.085	0.000	0.981
4	T	E	0.020	0.000	0.996
5	F	E	0.029	0.001	0.991
6	Y	E	0.080	0.005	0.969
7	E	E	0.340	0.010	0.770
8	D	C	0.944	0.005	0.089
9	R	C	0.907	0.029	0.030
10	G	C	0.962	0.020	0.006
11	F	C	0.885	0.019	0.070
12	Q	C	0.523	0.012	0.490
13	G	E	0.243	0.007	0.576
14	R	E	0.306	0.019	0.531
15	H	E	0.174	0.016	0.811
16	Y	E	0.131	0.056	0.756
17	E	E	0.236	0.054	0.551
18	C	H	0.223	0.342	0.231
19	S	C	0.454	0.216	0.134
20	S	C	0.582	0.206	0.066

Figure 2. The input vectors called “secondary structure profiles”.

AA	Hydropathy index	Scaled index
A	1.8	1.8 / 4.5
C	2.5	2.5 / 4.5
D	-3.5	-3.5 / 4.5
E	-3.5	-3.5 / 4.5
F	2.8	2.8 / 4.5
G	-0.4	-0.4 / 4.5
H	-3.2	-3.2 / 4.5
I	4.5	4.5 / 4.5
K	-3.9	-3.9 / 4.5
L	3.8	3.8 / 4.5
M	1.9	1.9 / 4.5
N	-3.5	-3.5 / 4.5
P	1.6	1.6 / 4.5
Q	-3.5	-3.5 / 4.5
R	-4.5	-4.5 / 4.5
S	-0.8	-0.8 / 4.5
T	-0.7	-0.7 / 4.5
V	4.2	4.2 / 4.5
W	-0.9	-0.9 / 4.5
Y	-1.3	-1.3 / 4.5

1	T	-0.7 / 4.5
2	S	-0.8 / 4.5
3	P	1.6 / 4.5
4	Q	-3.5 / 4.5
5	R	-4.5 / 4.5
6	E	-3.5 / 4.5
7	A	1.8 / 4.5
8	T	-0.7 / 4.5
9	C	2.5 / 4.5
10	T	-0.7 / 4.5
11	S	-0.8 / 4.5
12	E	-3.5 / 4.5
13	V	4.2 / 4.5
14	S	-0.8 / 4.5
15	G	-0.4 / 4.5
16	C	2.5 / 4.5
17	P	1.6 / 4.5
18	K	-3.9 / 4.5
19	I	4.5 / 4.5
20	Y	-1.3 / 4.5

15*1 scaled inputs

Figure 3. The input vectors called "hydropathy indexes".

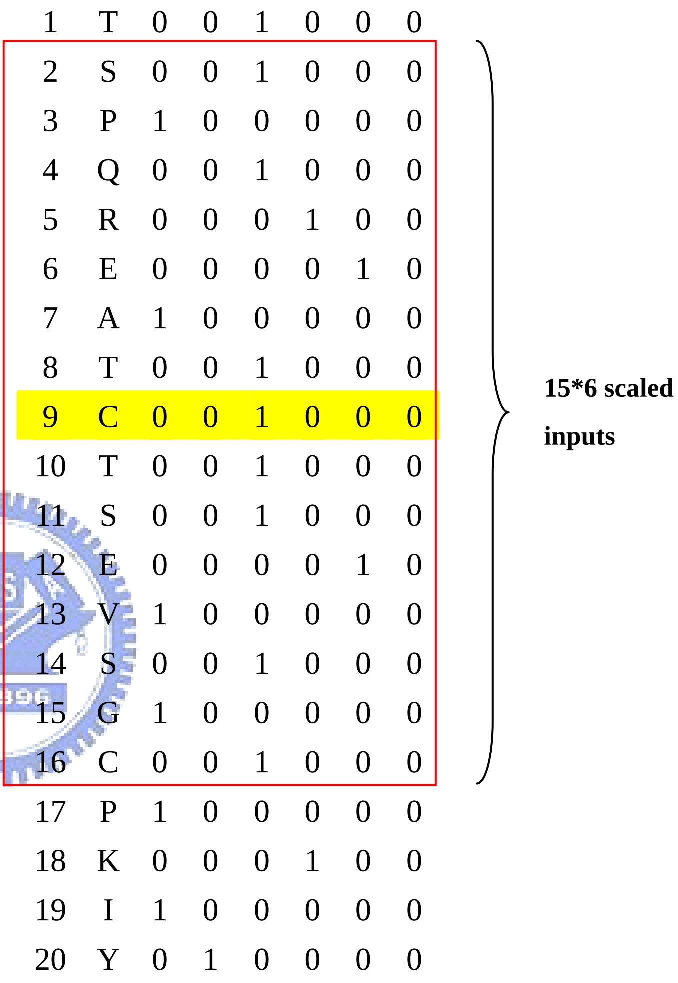
AA	ID
A	1
C	2
D	3
E	4
F	5
G	6
H	7
I	8
K	9
L	10
M	11
N	12
P	13
Q	14
R	15
S	16
T	17
V	18
W	19
Y	20

1	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2	S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
6	E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
9	C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
18	K	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
19	I	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
20	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

15*20 scaled inputs

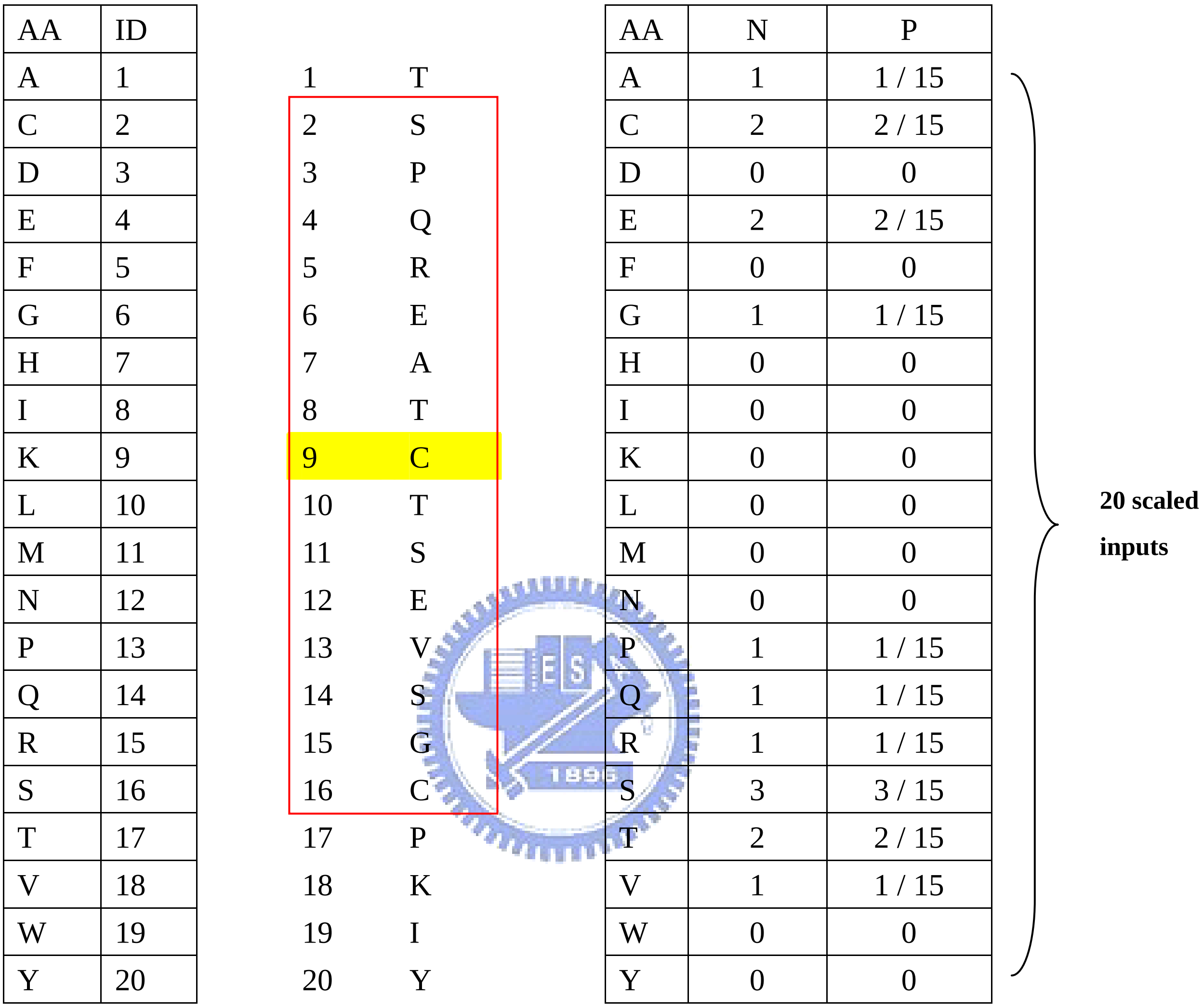
Figure 4. The input vectors called “amino acid in binary coding”.

Class	ID	AA
Nonpolar, aliphatic R groups	1	G
	1	A
	1	P
	1	V
	1	L
	1	I
	1	M
Aromatic R groups	2	F
	2	Y
	2	W
Polar, uncharged R groups	3	S
	3	T
	3	C
	3	N
	3	Q
Positively charged R groups	4	K
	4	H
	4	R
Negatively charged R groups	5	D
	5	E
Others	6	



1	T	0	0	1	0	0	0
2	S	0	0	1	0	0	0
3	P	1	0	0	0	0	0
4	Q	0	0	1	0	0	0
5	R	0	0	0	1	0	0
6	E	0	0	0	0	1	0
7	A	1	0	0	0	0	0
8	T	0	0	1	0	0	0
9	C	0	0	1	0	0	0
10	T	0	0	1	0	0	0
11	S	0	0	1	0	0	0
12	E	0	0	0	0	1	0
13	V	1	0	0	0	0	0
14	S	0	0	1	0	0	0
15	G	1	0	0	0	0	0
16	C	0	0	1	0	0	0
17	P	1	0	0	0	0	0
18	K	0	0	0	1	0	0
19	I	1	0	0	0	0	0
20	Y	0	1	0	0	0	0

Figure 5. The input vectors called “amino acid properties in binary coding”.



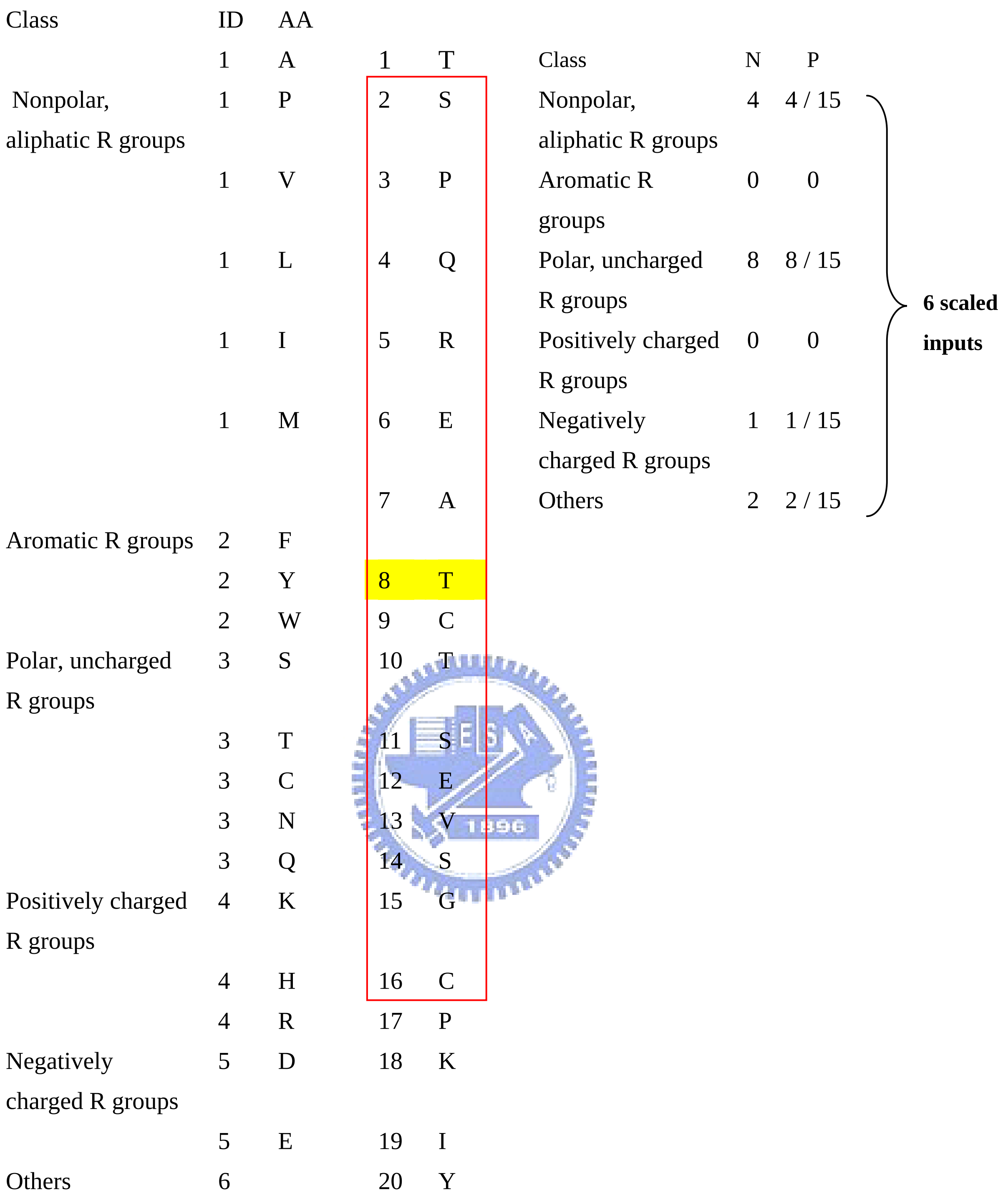
AA	ID
A	1
C	2
D	3
E	4
F	5
G	6
H	7
I	8
K	9
L	10
M	11
N	12
P	13
Q	14
R	15
S	16
T	17
V	18
W	19
Y	20

1	T
2	S
3	P
4	Q
5	R
6	E
7	A
8	T
9	C
10	T
11	S
12	E
13	V
14	S
15	G
16	C
17	P
18	K
19	I
20	Y

AA	N	P
A	1	1 / 15
C	2	2 / 15
D	0	0
E	2	2 / 15
F	0	0
G	1	1 / 15
H	0	0
I	0	0
K	0	0
L	0	0
M	0	0
N	0	0
P	1	1 / 15
Q	1	1 / 15
R	1	1 / 15
S	3	3 / 15
T	2	2 / 15
V	1	1 / 15
W	0	0
Y	0	0

20 scaled inputs

Figure 6. The input vectors called “local amino acid composition”.



Class	ID	AA
	1	A
Nonpolar, aliphatic R groups	1	P
	1	V
	1	L
	1	I
	1	M
Aromatic R groups	2	F
	2	Y
	2	W
Polar, uncharged R groups	3	S
	3	T
	3	C
	3	N
	3	Q
Positively charged R groups	4	K
	4	H
	4	R
Negatively charged R groups	5	D
	5	E
Others	6	

1	T
2	S
3	P
4	Q
5	R
6	E
7	A
8	T
9	C
10	T
11	S
12	E
13	V
14	S
15	G
16	C
17	P
18	K
19	I
20	Y

Class	N	P
Nonpolar, aliphatic R groups	4	4 / 15
Aromatic R groups	0	0
Polar, uncharged R groups	8	8 / 15
Positively charged R groups	0	0
Negatively charged R groups	1	1 / 15
Others	2	2 / 15

Figure 7. The input vectors called “local amino acid properties composition”.

AA	MaxACC	Scaled MaxACC
A	117.2	117.2 / 270.7
C	142.0	142.0 / 270.7
D	169.8	169.8 / 270.7
E	202.0	202.0 / 270.7
F	233.0	233.0 / 270.7
G	87.9	87.9 / 270.7
H	195.4	195.4 / 270.7
I	182.1	182.1 / 270.7
K	214.2	214.2 / 270.7
L	176.2	176.2 / 270.7
M	204.0	204.0 / 270.7
N	169.6	169.6 / 270.7
P	148.9	148.9 / 270.7
Q	199.4	199.4 / 270.7
R	242.9	242.9 / 270.7
S	135.2	135.2 / 270.7
T	152.8	152.8 / 270.7
V	162.4	162.4 / 270.7
W	270.7	270.7/ 270.7
Y	253.8	253.8/ 270.7

1	T	152.8 / 270.7
2	S	135.2 / 270.7
3	P	148.9 / 270.7
4	Q	199.4 / 270.7
5	R	242.9 / 270.7
6	E	202.0 / 270.7
7	A	117.2 / 270.7
8	T	152.8 / 270.7
9	C	142.0 / 270.7
10	T	152.8 / 270.7
11	S	135.2 / 270.7
12	E	202.0 / 270.7
13	V	162.4 / 270.7
14	S	135.2 / 270.7
15	G	87.9 / 270.7
16	C	142.0 / 270.7
17	P	148.9 / 270.7
18	K	214.2 / 270.7
19	I	182.1 / 270.7
20	Y	253.8/ 270.7

15 scaled inputs

Figure 8. The input vectors called "residue size".

AA	ID
A	1
C	2
D	3
E	4
F	5
G	6
H	7
I	8
K	9
L	10
M	11
N	12
P	13
Q	14
R	15
S	16
T	17
V	18
W	19
Y	20

1	T
2	S
3	P
4	Q
5	R
6	E
7	A
8	T
9	C
10	T
11	S
12	E
13	V
14	S
15	G
16	C
17	P
18	K
19	I
20	Y

AA	N	P
A	1	1 / 20
C	2	2 / 20
D	0	0
E	2	2 / 20
F	0	0
G	1	1 / 20
H	0	0
I	1	1 / 20
K	1	1 / 20
L	0	0
M	0	0
N	0	0
P	2	2 / 20
Q	1	1 / 20
R	1	1 / 20
S	3	3 / 20
T	3	3 / 20
V	1	1 / 20
W	0	0
Y	1	1 / 20

20 scaled inputs

Figure 9. The input vectors called “global amino acid composition”.

Class	ID	AA
Nonpolar, aliphatic R groups	1	G
	1	A
	1	P
	1	V
	1	L
	1	I
	1	M
Aromatic R groups	2	F
	2	Y
	2	W
Polar, uncharged R groups	3	S
	3	T
	3	C
	3	N
	3	Q
Positively charged R groups	4	K
	4	H
	4	R
Negatively charged R groups	5	D
	5	E
Others	6	

1	T
2	S
3	P
4	Q
5	R
6	E
7	A
8	T
9	C
10	T
11	S
12	E
13	V
14	S
15	G
16	C
17	P
18	K
19	I
20	Y

Class	N	P
Nonpolar, aliphatic R groups	6	6 / 20
Aromatic R groups	1	1 / 20
Polar, uncharged R groups	9	9 / 20
Positively charged R groups	1	1 / 20
Negatively charged R groups	1	1 / 20
Others	2	2 / 20

6 caled inputs

Figure 10. The input vectors called “global amino acid properties composition”.

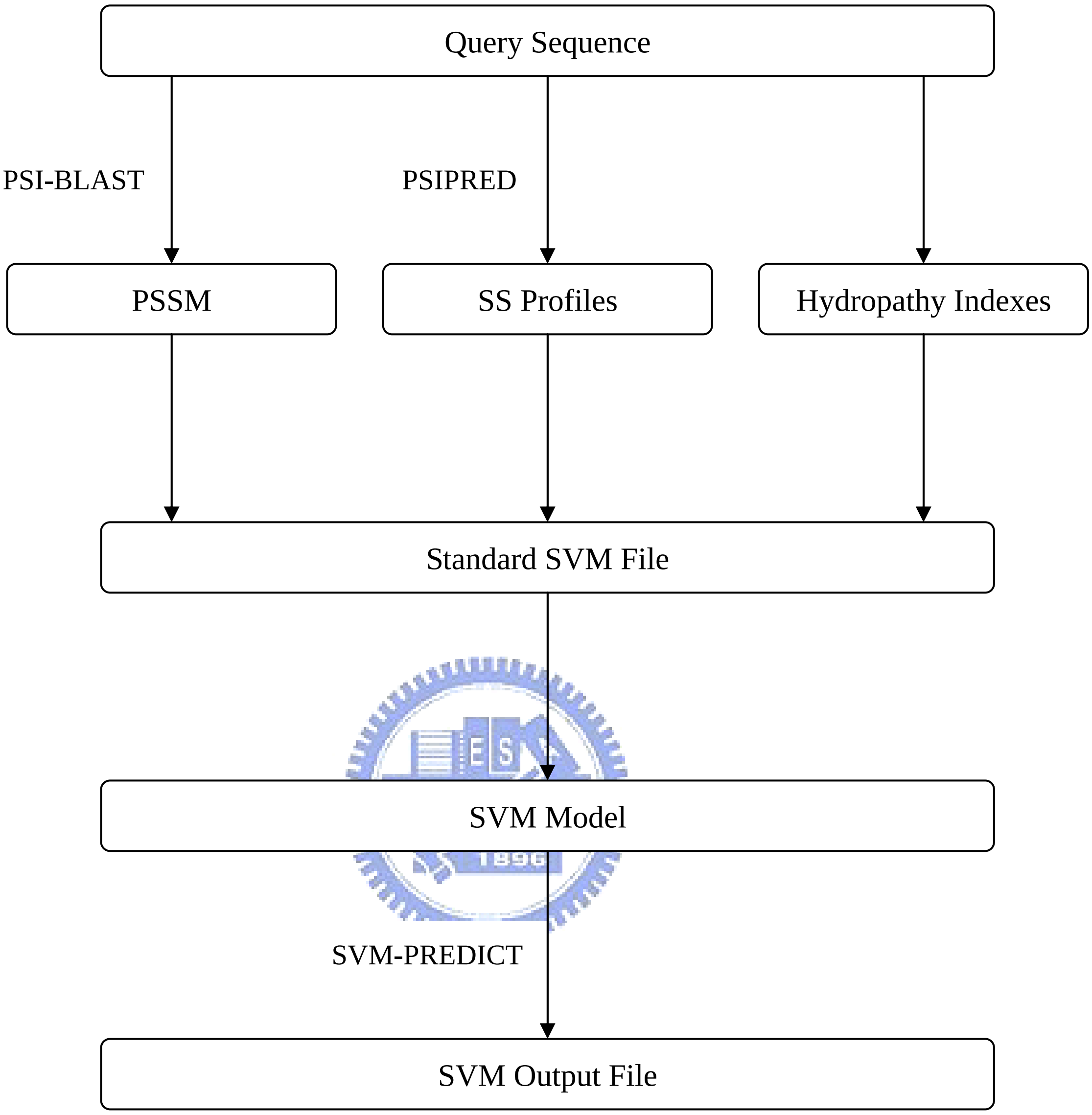
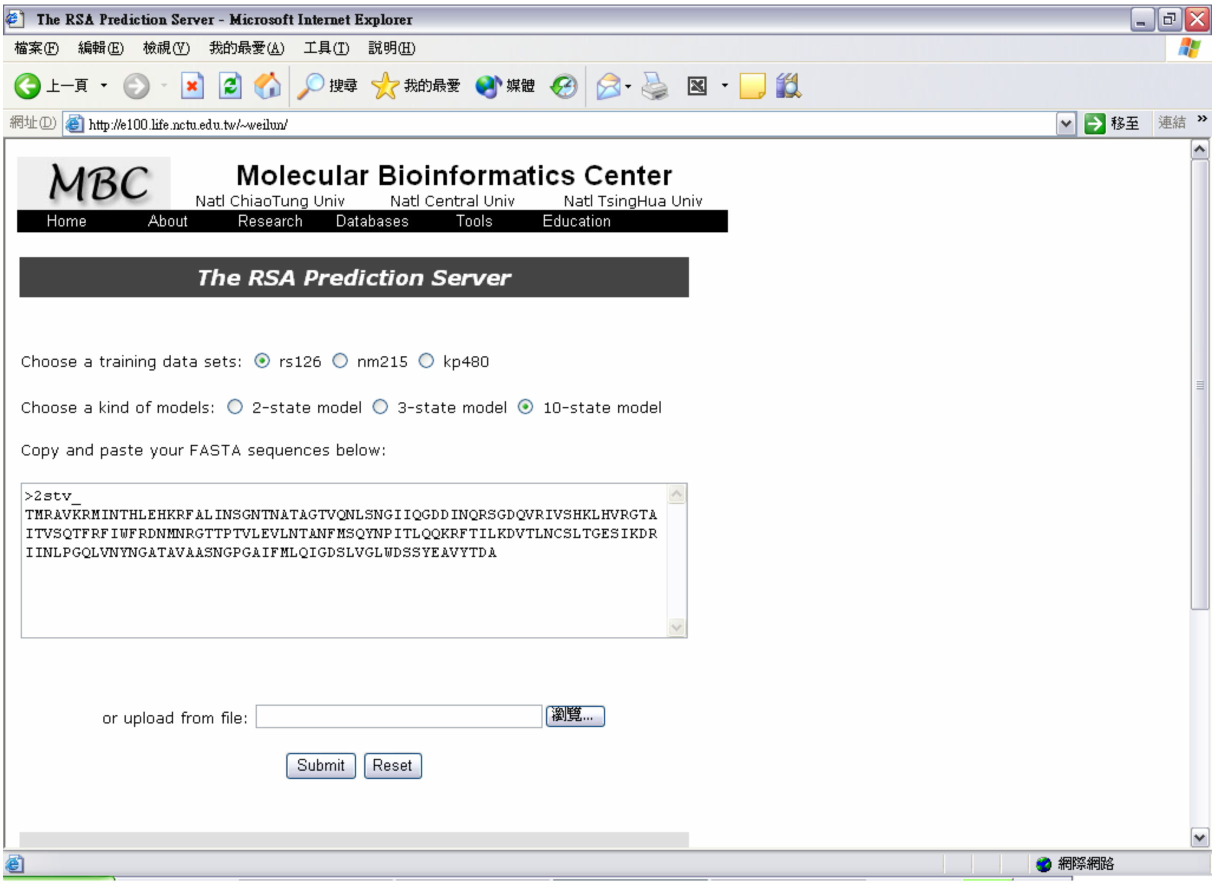


Figure 11. The figure shows the components of SAS prediction server, and most of the connection was built using the Perl language and CGI package.



The RSA Prediction Server - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D) http://e100.life.nctu.edu.tw/~weilun/
移至 連結
MBC
Molecular Bioinformatics Center
Natl ChiaoTung Univ Natl Central Univ Natl TsingHua Univ
Home About Research Databases Tools Education
The RSA Prediction Server

Choose a training data sets: ● rs126 ○ nm215 ○ kp480

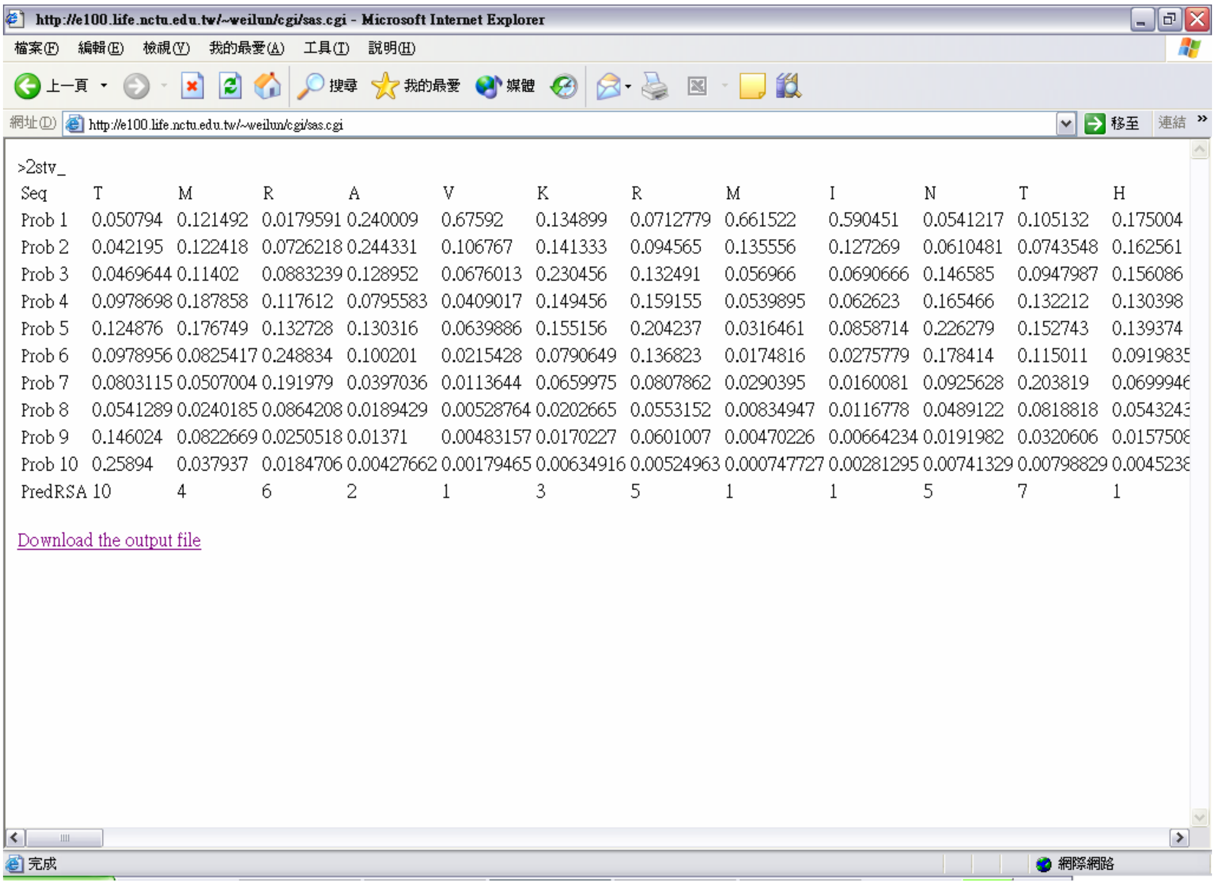
Choose a kind of models: ○ 2-state model ○ 3-state model ● 10-state model

Copy and paste your FASTA sequences below:

```
>2stv_
TMRAVKRMINTHLEHKRFALINSGNTNATAGTVQNLSNGIIQGDDINQRSGDQVRIVSHKLHVRGTA
ITVSQTFRFIWFRDNMNRGTTPTVLEVLNTANFMSQYNPITLQQKRFTILKDVTLNCSLTGESIKDR
IINLPGQLVNYNGATAVAASNGPGAIFMLQIGDSLVGLWDSSYEAVYTDA
```

or upload from file:

Figure 12. The World Wide Web interface of the RSA prediction server.


http://e100.life.nctu.edu.tw/~weilun/cgi/sas.cgi - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D) http://e100.life.nctu.edu.tw/~weilun/cgi/sas.cgi
移至
連結
完成
網際網路

```
>2stv_
Seq      T         M         R         A          V          K          R          M           I          N          T          H
Prob 1   0.050794  0.121492  0.0179591 0.240009   0.67592    0.134899   0.0712779  0.661522    0.590451   0.0541217  0.105132   0.175004
Prob 2   0.042195  0.122418  0.0726218 0.244331   0.106767   0.141333   0.094565   0.135556    0.127269   0.0610481  0.0743548  0.162561
Prob 3   0.0469644 0.11402   0.0883239 0.128952   0.0676013  0.230456   0.132491   0.056966    0.0690666  0.146585   0.0947987  0.156086
Prob 4   0.0978698 0.187858  0.117612  0.0795583  0.0409017  0.149456   0.159155   0.0539895   0.062623   0.165466   0.132212   0.130398
Prob 5   0.124876  0.176749  0.132728  0.130316   0.0639886  0.155156   0.204237   0.0316461   0.0858714  0.226279   0.152743   0.139374
Prob 6   0.0978956 0.0825417 0.248834  0.100201   0.0215428  0.0790649  0.136823   0.0174816   0.0275779  0.178414   0.115011   0.0919835
Prob 7   0.0803115 0.0507004 0.191979  0.0397036  0.0113644  0.0659975  0.0807862  0.0290395   0.0160081  0.0925628  0.203819   0.0699946
Prob 8   0.0541289 0.0240185 0.0864208 0.0189429  0.00528764 0.0202665  0.0553152  0.00834947  0.0116778  0.0489122  0.0818818  0.0543243
Prob 9   0.146024  0.0822669 0.0250518 0.01371    0.00483157 0.0170227  0.0601007  0.00470226  0.00664234 0.0191982  0.0320606  0.0157508
Prob 10  0.25894   0.037937  0.0184706 0.00427662 0.00179465 0.00634916 0.00524963 0.000747727 0.00281295 0.00741329 0.00798829 0.0045238
PredRSA  10        4         6         2          1          3          5          1           1          5          7          1
```

Download the output file

*Prob X: probability of the residue belongs to class X

*PredRSA: the predicted relative class

Figure 13. The standard output webpage of prediction.

Tables

Table 1. Maximum value of solvent accessible surface areas of each kind of residue calculated by MOLSIM program

Amino acid	MaxACC
A	117.2
B	160.0
C	142.0
D	169.8
E	202.0
F	233.0
G	87.9
H	195.4
I	182.1
K	214.2
L	176.2
M	204.0
N	169.6
P	148.9
Q	199.4
R	242.9
S	135.2
T	152.8
V	162.4
W	270.7
X	180.0
Y	253.8
Z	196.0

Table 2. The RS126 dataset used in this study

256b_a	8abp_	1acx_	3ait_	2alp_
9api_b	1cyo_	1bds_	1bmv_2	4bp2_
7cat_a	1cc5_	1cdh_	3cla_	4cms_
6cpa_	4cpv_	1cse_i	2cyp_	1eca_
3ebx_	1etu_	1fdl_h	1fkf_	1iqz_
2fox_	2gbp_	1gdl_o	2gn5_	4gr1_
6hir_	3hmg_b	5hvp_a	3icb_	1il8_a
1l58_	5ldh_	2lhb_	2ltn_a	5lyz_
2mev_4	1ovo_a	1paz_	2pcy_	3pgm_
1pyp_	2mhu_	1ppt_	1rhd_	4rhv_3
3rnt_	2rsp_a	1s01_	4sgb_i	2sns_
2stv_	1tgs_i	6tmn_e	1tnf_a	2tsc_a
2utg_a	2wrp_r	1bks_b	1prc_c	1prc_l
2aat_	6acn_	8adh_	2ak3_a	9api_a
1azu_	1bbp_a	1bmv_1	3blm_	2cab_
1cbh_	2ccy_a	1cdt_a	3cln_	4cpa_i
6cpp_	1crn_	6cts_	5cyt_r	6dfr_
5er2_e	1fc2_c	1dur_	1fnd_	1fxi_a
1g6n_a	1a45_	2gls_a	1gp1_a	1hip_
3hmg_a	2hmz_a	2i1b_	7icd_	9ins_b
1lap_	1gdj_	1lmb_3	2ltn_b	1mcp_l
2or1_l	2pab_a	9pap_	4pfk_	2phh_
1r09_2	1mrt_	1rbp_	4rhv_1	4rhv_4
7rsa_	4rxn_	3sdh_a	1sh1_	2sod_b
2tgp_i	3tim_a	2tmv_p	4ts1_a	1ubq_
9wga_a	1bks_a	4xia_a	1prc_h	1prc_m

Table 3. The NM215 dataset used in this study

119l_	2end_	1thv_	1pex_	1hsba	1csn_	1dela
1bnca	3chy_	1xgsa	1rro_	1mbd_	1fjma	1gai_
1dktb	1abrb	2hhma	1tml_	1php_	1ido_	1itg_
1ggga	1cem_	3mdda	1xyza	1sbp_	1mml_	1nbab
1knya	1dxy_	1amm_	2liv_	1ubi_	1pmi_	1poc_
1ofga	1gpc_	1cfya	3sdha	1ytw_	1slua	1sra_
1pud_	1kuh_	1ede_	1atla	2pgd_	1uxy_	1vhra
1svpa	1ospo	1gtma	1chma	6gsva	2arca	1m85_
1vls_	1ra9_	1lki_	1erv_	1bbpa	2pia_	2rspb
2chsa	1tfe_	1pdo_	1hgxa	1csee	8atcb	1bib_
2sns_	1who_	1rgs_	1mai_	1ezm_	1bfg_	1dfji
153l_	2gdm_	1thx_	1pgs_	1htp_	1cyx_	1gcb_
1btma	3cox_	1xnb_	1rsy_	1mkaa	1ftpa	1jkw_
1dkza	1afra	2hpda	1tupc	1pioa	1ifc_	1nox_
1gnd_	1ceo_	3minb	1yasa	1sftb	1mola	1pot_
1kpta	1ecea	1amp_	2mtac	1uby_	1pne_	1std_
1onra	1gpl_	1chd_	5p21_	256ba	1smea	1vid_
1pyta	1lba_	1edg_	1atna	2phla	1vcaa	2cba_
1tadc	1pbc_	1hava	1cmke	6pfka	2ayh_	2scpa
1wba_	1rcf_	1lkka	1esc_	1bdo_	2pspa	1bmfg
2ctc_	1tfr_	1pea_	1hlb_	1csga	1bgc_	1dhr_
2tysa	1bksb	1rnl_	1maz_	1fds_	1deaa	1gdoa
1aba_	2hft_	1tib_	1phe_	1idaa	1fua_	1knb_
1btn_	3grs_	1xvaa	1rvaa	1mlda	1irk_	1noza
1dosa	1afwa	2i1b_	1tys_	1plc_	1nar_	1ppn_
1gotb	1cewi	3nll_	1ysc_	1sig_	1poa_	1stfi
1kte_	1ecpa	1aoca	2naca	1udii	1smpi	1vin_
1opr_	1gsa_	1chka	5ptp_	2abk_	1vhh_	2ccya
1qapa	1lcl_	1edt_	1axn_	2phy_	2bbvc	2sil_
2tdx_	1pda_	1hfc_	1cnv_	7rsa_	2rn2_	
1whi_	1rec_	1ltsa	1exnb	1beo_	1bhmb	

Table 4. The KP480 dataset used in this study

154l-1-AUTO.1
1aazb-1-DOMAK
1acx
1add-1-AS
1adeb-2-AUTO.1
1ahb-2-GJB
1alkb-1-AS
1amp-1-AS
1aorb-1-AS
1aorb-3-AS
1aozb-1-AS
1aozb-2-AS
1aozb-3-AS
1asw-1-AUTO.1
1avhb-3-AS
1avhb-4-AS
1ayab-1-GJB
1azu
1bam-1-AS
1bbpa
1bcx-1-DOMAK
1bdo-1-AS
1bet-1-DOMAK
1bfg-1-DOMAK
1bmv1
1bmv2
1bncb-1-AS
1bncb-3-AS
1bncb-4-AS
1bovb-1-DOMAK
1brse-1-DOMAK
1bsdb-1-DOMAK
1cbg-1-AS
1cbh
1cc5
1cdta
1cei-1-GJB
1celb-1-AUTO.1
1cem-1-GJB
1cewi-1-DOMAK
1cfb-1-AS
1cgu-2-GJB
1cgu-3-GJB
1cgu-4-GJB
2chbe-1-DOMAK
1chd-1-AS
1chkb-2-AUTO.1
1chmb-1-DOMAK
1cksc-1-AUTO.1
1clc-1-AS.1
1clc-2-AS.1
1clc-3-AS.1
1cnsb-1-AUTO.1
1colb-1-DOMAK
1comc-1-DOMAK
1cpcl-1-DOMAK
1cpn-1-DOMAK
1cqa-1-AUTO.1
1crn
1csei
1csmb-1-AUTO.1
1ctf-1-DOMAK
2cthb-1-DOMAK
1ctm-2-DOMAK
1ctn-1-AS.1
1ctn-3-AS.1
1ctu-1-AUTO.1
1ctu-2-AUTO.1
1eu1a-4-AUTO.1
1cyx-1-AUTO.1
1daab-1-AS
1daab-2-AS
1dar-3-AS
1dfji-1-AUTO.1
1dih-2-AS
1dik-1-AS.1
1dik-2-AS.1
1dik-3-AS.1
1dik-4-AS.1
1din-1-AS
1dkza-1-JAC
1dlc-1-AS.1
1dlc-3-AS.1
1dnpb-1-AUTO.1
1dnpb-2-AUTO.1
1dpgb-1-AUTO.1
1dpgb-2-AUTO.1
1dsbb-2-AUTO.1
1dts-1-AUTO.1
1dupa-1-AS
1dynb-1-AUTO.1
1eca
1eceb-1-AUTO.1
1ecl-1-AS
1ecl-4-AS
1ecpf-1-AUTO.1
1edd-1-DOMAK
1edmc-1-AUTO.1
1eft-3-DOMAK
1efud-2-AUTO.1
1epbb-1-DOMAK
1ese-1-AUTO.1
1esl-1-GJB
1etu
1euu-2-JAC
1fbab-1-DOMAK
1fbl-1-AS
1fc2c
1fdlh
1fdt-1-AS
1dur
1find-1-AUTO.1
1find-2-AUTO.1
1fjmb-2-AS
1fkf
1fnd
1fua-1-AUTO.1
1fuqb-1-AUTO.1
1fuqb-2-AUTO.1
1fuqb-3-AUTO.1
1fxia
1gal-2-AS
1gal-3-AS
1gcb-2-AS
1gcmc-1-AUTO.1
1gd1o
1gdj
2gep-2-AS
2gep-3-AS
1gflb-1-AS
1ghsb-1-GJB
1gky-2-AS
1gln-2-AS
1gln-4-AS
1gmpb-1-DOMAK
1gnd-2-JAC
1gog-1-AS.1
1gog-2-AS.1
1gog-3-AS.1
1gp1a
1gp2g-2-AS
1gpc-1-AS
1gpmd-4-AS
1gpmd-5-AS
1grj-1-AS
1grj-2-AS
1gtmc-2-AUTO.1
1gtqb-1-AUTO.1
1gym-1-AUTO.1
1han-1-AUTO.1
1han-2-AUTO.1
1hcgb-1-AS
1hcra-1-DOMAK
1hip
1hiws-1-AS
1hjrd-1-AUTO.1
1hmpb-1-AUTO.1
1hnf-1-AS
1hnf-2-AS
1horb-1-AUTO.1

Table 4. The KP480 dataset used in this study (continued)

1hplb-1-AS
1hplb-2-AS
1hslb-2-DOMAK
1htrp-1-AS
1hvq-1-AUTO.1
1hxn-1-AS
1hyp-1-DOMAK
1ignb-2-GJB
1il8a
1ilk-1-AS
1ilk-2-AS
1inp-1-AS.1
1inp-2-AS.1
1irk-1-AS
1irk-2-AS
1isab-1-GJB
1isab-2-GJB
1isub-1-DOMAK
1jud-1-GJB
1kimb-1-AUTO.1
1knb-1-AS
1krca-1-AUTO.1
1krcb-1-AS
1kte-1-AS
1ktq-1-AUTO.1
1kuh-1-AS
1l58
1lap
1latb-1-AUTO.1
1lba-1-DOMAK
1lbu-1-AS
1lbu-2-AS
1lehb-3-AS
1lib-1-DOMAK
1lis-1-DOMAK
1lki-1-AS
1lmb3
1lpba-1-DOMAK
1lpe-1-DOMAK
1mai-1-JAC
1masb-1-AUTO.1
1mdam-1-DOMAK
1mdta-1-AS
1mdta-2-AS
1mdta-3-AS
1mjc-1-DOMAK
1mla-2-AS.1
1mmoh-1-AS
1mns-2-AS
1mof-1-AS
1mrrb-1-DOMAK
1mrt
1mspb-1-AS
1nal4-1-AUTO.1
1nar-1-DOMAK
1nbac-1-AS
1ncg-1-AUTO.2
1ndh-1-AS
1ndh-2-AS
1nfp-1-AS
1nga-2-AS.1
1nlkl-1-DOMAK
1nol-1-AUTO.2
1nox-1-GJB
1nozb-2-AUTO.1
1oacb-1-AS.1
1oacb-2-AS.1
1oacb-3-AS.1
1oacb-4-AS.1
1onrb-1-AUTO.1
1otgc-1-AS
1ovb-1-GJB
1ovoa
1oxy-3-AS
1oyc-1-AS
1paz
1pbp-2-DOMAK
1pbwb-1-AS
1pda-2-AS
1pda-3-AS
1pdnc-2-AS
1pdo-1-GJB
1pga-1-DOMAK
1pht-1-AUTO.1
1pii-2-DOMAK
1pkyc-2-AUTO.1
1pkyc-3-AUTO.1
1pmi-2-GJB
1pnmb-2-AS
1pnt-1-AS
1poc-1-DOMAK
1powb-1-DOMAK
1powb-2-DOMAK
1powb-3-DOMAK
1ppi-2-AS
1ppt
1ptr-1-AUTO.1
1ptx-1-AS
1pyp
1pyta-1-AS
1qbb-1-AUTO.1
1qbb-2-AUTO.1
1qbb-3-AUTO.1
1qbb-4-AUTO.1
1qrdb-1-AUTO.1
1r092
1rbp
1rec-1-DOMAK
1rec-2-DOMAK
1regy-1-AUTO.1
1reqc-2-AS
1rhd
1rhgc-1-DOMAK
1rie-1-GJB
1ris-1-DOMAK
1rlds-1-DOMAK
1rlr-1-JAC
1rlr-2-JAC
1rpo-1-AUTO.1
1rsy-1-AS
1rvvz-1-AUTO.1
1s01
1scud-1-AS
1scue-2-AS
1scue-3-AS
1seib-1-AUTO.1
1seib-2-AUTO.1
1sesa-2-AS
1sfe-1-AS
1sfe-2-AS
1sftb-2-AS
1sh1
1smnb-1-AUTO.1
1smpi-1-AS
1spbp-1-AS
1sra-1-AS
1srja-1-DOMAK
1stfi-1-DOMAK
1stme-1-AUTO.1
1svb-1-AS
1svb-2-AS
1tabi-1-DOMAK
1taq-2-AS
1tcra-2-GJB
1tfr-1-GJB
1thtb-1-AUTO.1
1thx-1-AUTO.1
1tie-1-DOMAK
1tif-1-AS
1tig-1-AUTO.1
1tml-1-AS
1tndb-2-DOMAK
1tnfa
1tplb-3-AS
1trb-2-AS
1trh-1-AS
1trkb-1-AS
1trkb-3-AS
1tsp-1-AS
2tssb-2-DOMAK

Table 4. The KP480 dataset used in this study (continued)

1tul-1-JAC
1tupc-1-AUTO.1
1ubq
1udh-1-AUTO.1
1umub-1-AS
1vcab-1-AUTO.1
1vcab-2-AUTO.1
1vcc-1-AS
1vhh-1-AS
1vhrb-2-AUTO.1
1vid-1-JAC
1vjs-3-GJB
1vmob-1-AS
1vnc-1-JAC
1vokb-1-AS
1vpt-1-JAC
1wapv-1-AUTO.1
1whi-1-AS
1bksa
1bksb
1xvab-1-GJB
1yptb-1-AUTO.1
1yrna-2-AS
1znbb-1-AS
1zymb-2-AUTO.1
256ba
2aaib-2-DOMAK
2aat
2abk-2-AS
2admb-1-AUTO.1
2admb-2-AUTO.1
2afnc-1-AUTO.1

2afnc-2-AUTO.1
2ak3a
2alp
2asr-1-DOMAK
2bat-1-GJB
2bltb-2-AUTO.1
2bopa-1-DOMAK
2cab
2ccya
2cmd-2-GJB
2cpo-1-AUTO.1
2cyp
2dkb-2-AS
2dln-1-AS
2dln-3-AS
2dnja-1-AS
2ebn-1-AS
2end-1-DOMAK
2fox
1iqz
2gbp
1a45
2glsa
2gn5
2gsq-2-AS
2hft-1-AS
2hft-2-AS
2hhmb-1-DOMAK
2hhmb-2-DOMAK
2hipb-1-DOMAK
2hmza
2hpr-1-DOMAK

2i1b
2ltna
2ltnb
2mev4
2mtac-1-AS
2nadb-2-AS.1
2npx-3-AS.1
2olba-2-AS
2olba-3-AS
2or1l
2paba
2pgd-1-AUTO.1
2pgd-2-AUTO.1
2phh
2phy-1-GJB
2polb-1-AS
2reb-1-DOMAK
2reb-2-DOMAK
2rsla-1-GJB
2rspa
2scpb-1-DOMAK
2sil-1-AS
2sns
2sodb
2spt-1-DOMAK
2spt-2-DOMAK
2tgi-1-DOMAK
2tgpi
2tmdb-3-AS
2tmvp
2trt-1-AUTO.1
2tsca

2utga
2wrpr
2yhx-3-DOMAK
3ait
1cyo
4bcl-1-DOMAK
3blm
3cd4
3chy-1-DOMAK
3cla
3cln
3cox-1-AS.1
3cox-2-AS.1
3ecab-1-AS
3ecab-2-AS
1g6na
3hmga
3hmgb
3icb
3inkd-1-DOMAK
3mddb-1-AS
3mddb-2-AS
3mddb-3-AS
3pgk-2-AS
3pgm
3pmgb-1-AS
3pmgb-2-AS
3pmgb-3-AS
3pmgb-4-AS
3rnt
3tima
4bp2

4fisb-1-DOMAK
4gr1
4pfk
4rhv1
4rhv3
4rhv4
4rxn
4sdha
4sgbi
4ts1a
4xiaa
5cytr
5er2e
5ldh
5lyz
5sici-1-DOMAK
6acn
6cpa
6cpp
6cts
6dfr
6hir
6tmne
7cata
7icd
7rsa
821p-1-DOMAK
8adh
9apia
9apib
9pap
9wgaa

Table 5. The hydropathy index of the standard amino acids.

Amino acid	Hydropathy index
A	1.8
C	2.5
D	-3.5
E	-3.5
F	2.8
G	-0.4
H	-3.2
I	4.5
K	-3.9
L	3.8
M	1.9
N	-3.5
P	1.6
Q	-3.5
R	-4.5
S	-0.8
T	-0.7
V	4.2
W	-0.9
Y	-1.3

Table 6. The classification of the standard amino acids by R groups.

Class	Amino acid
Nonpolar, aliphatic R groups	G
	A
	P
	V
	L
	I
	M
Aromatic R groups	F
	Y
	W
Polar, uncharged R groups	S
	T
	C
	N
	Q
Positively charged R groups	K
	H
	R
Negatively charged R groups	D
	E

Table 7. Prediction accuracies and MCC values using individual local descriptors with RS126 datasets in 2-state prediction model.

Features	Threshold	Qtotal	Qb	Qe	MCC
PSSM	25	75.33	76.38	74.11	50.43
Hydropathy indexes	25	65.73	80.43	48.61	30.79
Secondary structure profile	25	67.80	68.93	66.46	35.34
Amino acid in binary coding	25	70.59	74.49	66.06	40.70
Amino acid properties in binary coding	25	67.34	69.51	64.81	34.31
Local amino acid composition	25	58.22	72.67	41.39	14.82
Local amino acid properties composition	25	58.04	75.85	37.30	14.27
Residue size	25	55.8	96.19	8.68	10.18

Table 8. Prediction accuracies and MCC values using comprehensive local descriptors with RS126 datasets in 2-state prediction model.

Features	Threshold	Qtotal	Qb	Qe	MCC
PSSM+H	25	76.45	78.96	73.52	52.57
PSSM+SS	25	75.98	79.06	72.40	51.60
H+SS	25	68.40	70.05	66.48	36.50
PSSM+H+SS	25	77.23	78.53	75.72	54.23
PSSM+H+SS+AA	25	77.78	79.44	75.84	55.29
PSSM+H+SS+PROP	25	77.40	79.15	75.36	54.53
PSSM+H+SS+AA+PROP	25	77.33	78.64	75.80	54.42
PSSM+H+SS+GAA	25	77.28	78.48	75.88	54.33
PSSM+H+SS+GPROP	25	77.15	78.29	75.82	54.07

*H: hydropathy indexes
*SS: secondary structure profile
*AA: amino acid in binary coding
*PROP: amino acid properties in binary coding
*GAA: Global amino acid composition
*GPROP: Global amino acid properties composition

Table 9. Prediction accuracies of the benchmark tests on the RS126, NM215 and KP480 datasets in 2-state prediction model and its comparison with other methods

Year	Method	State threshold					
		25%	20%	16%	5%	0%	9,36%
1994	PHDacc(RS126)[3]	-	-	74.8	-	86.0	57.5
2000	Jnet(CB480)[7]	76.2	-	-	79.8	86.5	-
2002	BRNNs (PDB-1086)[4]	77.2	77.5	-	81.2	86.5	-
2003	PP(RS-126)[23]	-	-	75.1	-	-	57.9
2003	NETASA(NM-215)[24]	70.3	-	-	74.6	87.9	-
2004	SVMpsi (RS126)[13]	76.8	-	77.8	79.8	86.2	59.6
2004	SVMpsi(KP480)[13]	78.7	-	80.7	82.4	87.4	64.5
2005	Fuzzy k-NN(RS-126)[14]	78.3	-	79.0	82.2	87.2	63.8
2005	SVR12(RA-603)[10]	76.7	76.8	-	76.5	-	-
2005	WE(2148 chains)[1]	-	80.0	-	-	-	-
	PSSM+H+SS (RS126)	77.2	77.6	77.9	81.4	88.4	60.2
	PSSM+H+SS+AA (RS126)	77.8	77.8	78.1	80.0	88.8	61.6
	PSSM+H+SS (NM215)	78.3	77.7	87.2	87.4	87.0	62.7
	PSSM+H+SS (KP480)	78.0		78.5	81.3	87.0	62.3

Table 10. MCC values of the benchmark tests on the RS126, NM215 and KP480 datasets in 2-state prediction model and its comparison with other methods

Year	Method	State threshold					
		25%	20%	16%	5%	0%	9,36%
1994	PHDacc(RS126)[3]	-	-	-	-	-	-
2000	Jnet(CB480)[7]	-	-	-	-	-	-
2002	BRNNs (PDB-1086)[4]	-	-	-	-	-	-
2003	PP(RS-126)[23]	-	-	48.5	-	-	53.0
2003	NETASA(NM-215)[24]	-	-	-	-	-	-
2004	SVMpsi (RS126)[13]	-	-	-	-	-	-
2004	SVMpsi(KP480)[13]	-	-	-	-	-	-
2005	Fuzzy k-NN(SCOP-229)[14]	55.4	-	56.0	54.1	43.1	56.0;19.9;50.8
2005	SVR12(RA-603)[10]	-	-	-	-	-	-
2005	WE(2148 chains)[1]	-	60.0	-	-	-	-
	PSSM+H+SS (RS126)	54.2	55.2	54.7	49.7	22.2	52.0;17.0;48.4
	PSSM+H+SS+AA (RS126)	55.3	55.6	55.2	45.7	31.7	53.2;19.3;50.8
	PSSM+H+SS (NM215)	56.0	56.5	34.2	36.4	38.3	56.3;20.9;50.7
	PSSM+H+SS (KP480)	55.1		56.6	53.2	35.8	55.2;19.9;50.9

Table 11. MAE and Pearson's "r" values of the benchmark tests on the RS126 and KP480 datasets in 10-state prediction model

Year	Method	Data Set	MAE	R	Accuracy
2003	RVP[16]	RS-126	19.4	0.477	71%(16%)
		Carugo-338	19.0	0.490	
		CB-502	18.8	0.482	
2004	SVR[25]	CB-502	18.5	0.520	
		YH-1277	17.0	0.617	74.6%(15%)
	SABLE[5]	RA-603	15.3-15.8	0.64-0.67	77%(15%)
2005	SVR12[18]	RA-603	15.8-16.6	0.62-0.63	75.8(15%)
	PSSM+H+SS	RS-126	15.2	0.506	72.8(20%)
		KP-480	14.1	0.537	74.4(20%)

參考文獻

1. Chen, H. and H.X. Zhou, *Prediction of solvent accessibility and sites of deleterious mutations from protein sequence.* Nucleic Acids Res, 2005. **33**(10): p. 3193-9.
2. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.
3. Rost, B. and C. Sander, *Conservation and prediction of solvent accessibility in protein families.* Proteins, 1994. **20**(3): p. 216-26.
4. Pollastri, G., et al., *Prediction of coordination number and relative solvent accessibility in proteins.* Proteins, 2002. **47**(2): p. 142-53.
5. Adamczak, R., A. Porollo, and J. Meller, *Accurate prediction of solvent accessibility using neural networks-based regression.* Proteins, 2004. **56**(4): p. 753-67.
6. Ahmad, S. and M.M. Gromiha, *Design and training of a neural network for predicting the solvent accessibility of proteins.* J Comput Chem, 2003. **24**(11): p. 1313-20.
7. Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction.* Proteins, 2000. **40**(3): p. 502-11.
8. Thompson, M.J. and R.A. Goldstein, *Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes.* Proteins, 1996. **25**(1): p. 38-47.
9. Li, X. and X.M. Pan, *New method for accurate prediction of solvent accessibility from protein sequence.* Proteins, 2001. **42**(1): p. 1-5.
10. Wagner, M., et al., *Linear regression models for solvent accessibility prediction in proteins.* J Comput Biol, 2005. **12**(3): p. 355-69.
11. Naderi-Manesh, H., et al., *Prediction of protein surface accessibility with information theory.* Proteins, 2001. **42**(4): p. 452-9.
12. Yuan, Z., K. Burrage, and J.S. Mattick, *Prediction of protein solvent accessibility using support vector machines.* Proteins, 2002. **48**(3): p. 566-70.
13. Kim, H. and H. Park, *Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor.* Proteins, 2004. **54**(3): p. 557-62.
14. Sim, J., S.Y. Kim, and J. Lee, *Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method.* Bioinformatics, 2005. **21**(12): p. 2844-9.
15. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
16. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
17. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring*

matrices. J Mol Biol, 1999. **292**(2): p. 195-202.

18. Boser, B.E., Guyon,I.M. and Vapnik,V.N., *A training algorithm for optimal margin classifiers.* ACM Press, 1992: p. 144-152.

19. Hua, S. and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach.* J Mol Biol, 2001. **308**(2): p. 397-407.

20. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks.* Bioinformatics, 2001. **17**(4): p. 349-58.

21. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-14.

22. C.C. Chang, C.J.L., *Training nu-Support Vector Classifiers: Theory and Algorithms.* Neural Computation, 2001. **13(9)**: p. 2119-2147.

23. Gianese, G., F. Bossa, and S. Pascarella, *Improvement in prediction of solvent accessibility by probability profiles.* Protein Eng, 2003. **16**(12): p. 987-92.

24. Ahmad, S., M.M. Gromiha, and A. Sarai, *Real value prediction of solvent accessibility from amino acid sequence.* Proteins, 2003. **50**(4): p. 629-35.

25. Yuan, Z. and B. Huang, *Prediction of protein accessible surface areas by support vector regression.* Proteins, 2004. **57**(3): p. 558-64.