Predicting Protein Solvent Accessibility with Sequence, Evolutionary Information and Context-based Features

Ashraf Yaseen
Department of Mathematics & Computer Science
Central State University
Wilberforce, OH
ayaseen@centralstate.edu

Yaohang Li
Department of Computer Science
Old Dominion University
Norfolk, VA
yaohang@cs.odu.edu

***Abstract* — Solvent-accessible surface areas of residues in proteins are key factors in protein folding. Predicting solvent accessibility from protein sequences is significant for modeling the structural and functional characteristics of many proteins. In this work, we introduce an approach of enhancing solvent accessibility prediction accuracy. We derive pseudo-potentials, by considering high-order-inter-residue interactions, according to the amino acid environment around protein residues from large number of protein samples. These context-dependent pseudo-potentials are integrated as scores to estimate the favorability of a residue in solvent accessibility state. The context-based scores are then incorporated as features together with other sequence and evolutionary information to train 2-stage neural networks for solvent accessibility prediction. Our computational results have shown that the context-based scores are effective features to enhance the prediction accuracies of protein solvent accessibility. The 7-fold cross validated Q2 accuracy reached 80.76% when context-based scores are incorporated in the training process of the solvent accessibility predictor.**

***Index Terms* — Solvent accessibility, Context-based scores, Pseudo-potentials, Neural Networks**

I. INTRODUCTION

The solvent-accessible surface area, or accessibility, of a residue is the surface area of the residue that is exposed to solvent. The residue accessibility is a useful indicator to the residue's location, on the surface or in the core. Figure1 shows the surface area surrounding a protein segment.

Residue solvent accessibility is usually measured by rolling a spherical water molecule over a protein surface and summing the area that can be accessed by this molecule on each residue. To allow comparisons between the accessibility of the different amino acids in proteins, typically relative values are calculated as the ratio between the absolute solvent accessibility value and that in an extended tripeptide (Ala-X-Ala) conformation [1]; referred to as the percentage of maximally accessible area. Automated methods, like the DSSP program [2], can be used to calculate the absolute solvent accessibility values of proteins.

Residue solvent accessibility plays an important role in folding and enhancing proteins' thermodynamic and mechanical stability. The burial of residues at core (hydrophobic residues) is a major driving force for folding [3]. Moreover, the hydrophobic free energies are directly related to residues' solvent accessibilities, of both polar and nonpolar groups [4]. Furthermore, active sites of proteins are located on its surface. Hence, prediction of the surface residues is considered an important step in determining proteins functions [5].

Correctly predicting the solvent accessibility of residues can not only reduce the conformational space to aid modeling protein structures in three dimensions, but also help predict important protein functions.

A number of methods have been developed using different protein datasets and different computational methods, including neural networks [6-11], support vector machines [12, 13], nearest neighbor [14, 15], information theory [16], and Bayesian statistics [17]. In most of these methods, the prediction is performed in a discrete fashion, where predictors discriminate among a number of predefined levels or states of residues' exposure with predefined thresholds.

Predicting solvent accessibility using evolutionary information, revealed by multiple sequence alignments, led to a significant accuracy increase. Rost et al [18], Cuff et al [19], and Thompson et al [17] reported a two-state prediction accuracy of ~75% with 0.25 threshold. More recent prediction methods benefit from PSI-BLAST derived profiles to reach higher accuracies of ~78% in two-state prediction with 0.25 threshold, and an accuracy of ~64% in three-state prediction with 0.9 and 0.36 thresholds [9, 13-15].

Most of the current methods nowadays provide real value prediction, in addition to discrete-fashion prediction (in 2-state, 3-state, or more). The Pearson correlation coefficient (between the predicted and true values) reported in real value predictors is ~0.65 [14, 20].

Computational approaches for predicting solvent accessibility are mostly machine learning approaches, including statistical analysis, neural networks, SVM, hidden Markov Chains, etc. Features influencing the solvent accessibility of residues, such as

multiple sequence alignment, are encoded in the machine learning algorithms to improve prediction accuracy. Therefore, extracting and selecting "good" features are critical to the performance of the learning machines. Probably the most effective features for predicting the solvent accessibility state of a residue are the solvent accessibility states of its neighbors. For example, if both adjacent neighbors are buried, the middle residue is most likely to be buried, and vice versa. Unfortunately, using the true solvent accessibility states as features is not feasible since they cannot be known in advance. However, this inspires us that the favorability of a residue adopting a certain solvent accessibility state can be also an effective feature.

In this paper, we investigate the approaches of extracting context-based statistical scores, to measure the favorability of residues' solvent exposure in presence of its neighbors in sequence, from a large training sample set based on the mean-field potentials [21]. The fundamental idea is based on the fact that the residue's solvent accessibility exhibit strong local dependency. We derive statistics for singlets, doublets, and triplets in a sequence window from experimentally determined structures in PDB [22]. Then scores measuring the pseudo-energy of a residue adopting a certain accessibility state are calculated using potentials of mean force approach. These scores are then incorporated as features together with the multiple sequence alignment data to train neural networks for solvent accessibility prediction. 7-fold cross validations are performed. We apply our approach to predict solvent accessibility in 2-state. We also test our method on several commonly used protein benchmarks, including Manesh215 [1], Carugo338 [23], and CASP9 [24] targets. Lastly, we compare our method with a set of popular methods for solvent accessibility prediction, including NETASA [8], Sable [9], Netsurf [9], SPINE [6], ACCpro [11] and SANN [14].

II. METHOD

A. The Protein Data Sets

We use the protein chain dataset Cull16633 generated by the PISCES server [25] on 10/21/2011 to collect samples to generate context-based statistics. Cull16633 contains 16,633 chains with at most 50% sequence identity, 3.0A resolution cutoff, and 1.0 R-factor. For neural network training, we use the Cull7987 data set which includes 7,987 chains with at most 25% pair-wise sequence identity, 3.0A resolution cutoff, and 1.0 R-factor cutoff.

We use the PSI-BLAST program [26] to generate Position Specific Scoring Matrix (PSSM) data. Short chains less than 40 residues are eliminated, since the PSI-BLAST program is usually unable to generate profiles for very short sequences, and very large chains whose lengths are greater than 1000 residues. We also exclude residues with undetermined accessibility state from the training set.

The recent CASP9 targets as well as the public protein data sets Manesh215 and Carugo338, which are popularly employed as benchmarks for secondary structure predictions, are used to benchmark our method. Therefore, any sequences with greater than 25% similarity with the test benchmarks sequences are excluded from the Cull16633 and Cull7987 when context-based scores are generated and when the neural networks are trained. The resulting total number of proteins in Cull7987 and Cull16633 are 6,985 and 14,481, respectively.

The solvent accessibility values are determined by the DSSP program [2]. Relative values for residues' solvent accessibility are calculated as the ratio between the absolute solvent accessibility value and that in an extended tripeptide (Ala-X-Ala) conformation. Table 1 shows the extended state value of each amino acid reported by Ahmad et al. [1] and used in many prediction methods. A threshold of 0.25 is used to define the 2-state solvent accessibility (Buried when the relative solvent accessibility value is less than 0.25, and Exposed otherwise).

B. Context-based Statistics

It is well known that there exist general short range regularities in the primary structure of proteins [27]. Presumably, the neighboring residues have strong and probably deterministic influence to the chemical property of a residue in its accessibility to solvent [3].

Figure 2 shows the probability of residue K at position *i* in buried accessibility state with the neighboring residues at *i* - 1 and *i* + 1, *i* - 2 and *i* + 2, and *i* - 3 and *i* + 3 positions, respectively. One can notice that the residues separated by two residues in the middle still have strong influences on the state of the center residue.

In this work, we extract statistics of singlets (R_i), doublets (R_iR_{i+k}), and triplets ($R_iR_{i+k_1}R_{i+k_2}$) residues at different relative positions in protein sequences, which is further used to generate pseudo-potentials to be incorporated as new features in neural network training.

The statistics of singlets, doublets, and triplets represent estimations of the probabilities of residues adopting a specific solvent accessibility state when none, one, or two of their neighbors in context are taken into consideration, respectively.

Fig. 2. The probability of Lysine (K) as the middle residue of a triplet with neighboring residues at 1~3 positions away when adopting buried accessibility state. x, y, and - represent the left neighbor, right neighbor, and gap, respectively. The neighboring residues are ordered alphabetically.

The observed probabilities of the i^{th} residue R_i in a singlet (R_i), doublet (R_iR_{i+k}), and triplet ($R_iR_{i+k_1}R_{i+k_2}$) adopting a specific solvent accessibility C_i are respectively estimated by

$$P_{obs}(C_i|R_i) = \frac{N_{obs}(C_i, R_i)}{N_{obs}(R_i)},$$
$$P_{obs}(C_i|R_iR_{i+k}) = \frac{N_{obs}(C_i, R_iR_{i+k})}{N_{obs}(R_iR_{i+k})}, \text{and}$$
$$P_{obs}(C|R_iR_{i+k_1}R_{i+k_2}) = \frac{N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2})}{N_{obs}(R_iR_{i+k_1}R_{i+k_2})}.$$

Here $N_{obs}(C_i, R_i)$, $N_{obs}(C_i, R_iR_{i+k})$, and $N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2})$ are the weighted observed number of singlet (R_i), doublet (R_iR_{i+k}), and triplet ($R_iR_{i+k_1}R_{i+k_2}$) with R_i adopting conformation C_i (B, E for 2-state prediction) in the protein structure database (Cull16633). $N_{obs}(R_i)$, $N_{obs}(R_iR_{i+k})$, and $N_{obs}(R_iR_{i+k_1}R_{i+k_2})$ are the weighted observed number of singlets, doublets, and triplets

The frequency weights are obtained from the PSSM frequencies at each residue position in a protein sequence, which are generated by PSI-BLAST using three iterations of searching with e-value of 0.001 against the non-redundant (NR) database of protein sequences. The observed numbers are calculated as

$$N_{obs}(R_i) = \sum_{Protein} \sum_{j} PSSM_j(R_i),$$
$$N_{obs}(R_iR_{i+k}) = \sum_{Protein} \sum_{j} PSSM_j(R_i) * PSSM_j(R_{i+k}),$$
$$N_{obs}(R_iR_{i+k_1}R_{i+k_2}) = \sum_{Protein} \sum_{j} PSSM_j(R_i) * PSSM_j(R_{i+k_1}) * PSSM_j(R_{i+k_2}),$$
$$N_{obs}(C_i, R_i) = \sum_{Protein} \sum_{j}^{C_j=C_i} PSSM_j(R_i),$$
$$N_{obs}(C_i, R_iR_{i+k}) = \sum_{Protein} \sum_{\substack{j \\ C_j=C_i}} PSSM_j(R_i) * PSSM_j(R_{i+k}), \text{and}$$
$$N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2}) = \sum_{Protein} \sum_{\substack{j \\ C_j=C_i}} PSSM_j(R_i) * PSSM_j(R_{i+k_1}) * PSSM_j(R_{i+k_2}),$$

where $PSSM_j(R_i)$ is the PSSM frequency for residue type R_i at the jth position of a protein sequence.

C. Context-based Potential

The context-dependent pseudo-potentials are generated based on the potentials of mean force method. According to the inverse-Boltzmann theorem, we calculate the mean-force potential $U_{singlet}(R_i, C_i)$ for a singlet residue R_i adopting solvent accessibility state C_i,

$$U_{singlet}(C_i, R_i) = -RT\ln\frac{P_{obs}(C_i|R_i)}{P_{ref}(C_i|R_i)},$$

where R is the gas constant, T is the temperature, and $P_{ref}(C_i|R_i)$ is the referenced probability. In our method, we employ the conditional probability approach described in Samudrala and Moult to estimate the referenced probability by

$$P_{ref}(C_i|R_i) = \sum_{j}^{C_j=C_i} N_{obs}(C_j, R_j) \Big/ \sum_{j} N_{obs}(R_j).$$

Similarly, the mean-force potentials $U_{doublet}(C_i, R_iR_{i+k})$ and $U_{triplet}(C_i, R_iR_{i+k_1}R_{i+k_2})$ for residue adopting solvent accessibility are

$$U_{doublet}(C_i, R_iR_{i+k}) = -RT\ln\frac{P_{obs}(C_i|R_iR_{i+k})P_{ref}(C_i|R_i)}{P_{ref}(C_i|R_iR_{i+k})P_{obs}(C_i|R_i)}$$

and $U_{triplet}(C_i, R_iR_{i+k_1}R_{i+k_2}) = -RT\ln\frac{P_{obs}(C_i|R_iR_{i+k_1}R_{i+k_2})P_{ref}(C_i|R_iR_{i+k_2})P_{ref}(C_i|R_iR_{i+k_1})P_{obs}(C_i|R_i)}{P_{ref}(C_i|R_iR_{i+k_1}R_{i+k_2})P_{obs}(C_i|R_iR_{i+k_2})P_{obs}(C_i|R_iR_{i+k_1})P_{ref}(C_i|R_i)}$,

respectively with corresponding referenced probability

$$P_{ref}(C_i|R_iR_{i+k}) = \sum_{j}^{\substack{C_j=C_i \\ R_{j+k}=R_{i+k}}} N_{obs}(C_j, R_jR_{j+k}) \Big/ \sum_{j} N_{obs}(R_jR_{j+k}),$$

and

$$P_{ref}(C_i|R_iR_{i+k_1}R_{i+k_2}) = \sum_{\substack{j \\ R_{j+k_1}=R_{i+k_1} \\ R_{j+k_2}=R_{i+k_2}}}^{C_j=C_i} N_{obs}(C_j, R_jR_{j+k_1}R_{j+k_2}) \Big/ \sum_j N_{obs}(R_jR_{j+k_1}R_{j+k_2}).$$

Then, the context-dependent pseudo-potential for R_i under its amino acid environment is

$$U(C_i, \dots R_{i-1}R_iR_{i+1} \dots) = U_{singlet}(C_i, R_i) + \sum_k U_{doublet}(C_i, R_iR_{i+k}) + \sum_{k_1,k_2} U_{triplet}(C_i, R_iR_{i+k_1}R_{i+k_2}).$$

These context-dependent pseudo-potentials are used as context-based scores to encode in neural network training.

D. Neural Network Model

Our method incorporates two phases of neural network training. We adopt the standard feed-forward back-propagation neural network architecture. The first neural network phase is sequence-to-structure and the second phase is structure-to-structure training. The number of hidden nodes is 170 and 30 in the first and second networks, respectively.

In the sequence-to-structure training, a sliding window of 15 residues was selected, where each neural network is trained to predict the class of that residue in the middle of the window. Each residue is represented by 20 PSSM values and 1 extra value to indicate C- or N-terminals overlap. When the context-based scores are incorporated, additional 2 encoding values for each residue are needed. Overall, 360 input values are used to encode each residue in 2-state prediction.

After sequence-to-structure training, the next phase is to carry out a structure-to-structure training to eliminate impossible solvent accessibility predictions. Figure 3 depicts the encoding and neural network architecture for solvent accessibility prediction.

E. N-fold Cross validation

To have a reliable estimation of the prediction accuracy, we employ the 7-fold cross validation approach on the Cull data sets. The protein chains in the cull data sets are divided into 7 subsets with approximately the equal size. At each step, 5 subsets are used for neural network training while the other 2 are used separately for testing and validation. The process is repeated 7 times. The overall prediction accuracy is calculated as the average of the accuracies of the 7 folds.

III. RESULTS

We use Q_2 to measure the quality of our prediction method. Q_2 equals the total number of residues correctly predicted divided by the total number of residues. We also use Q_B and Q_E to measure the quality of predicting the buried state and the exposed state respectively.

Table II compares the prediction qualities of solvent accessibility with PSSM-only encoding and PSSM with context-based scores encoding after 7-fold cross validation. Compared to the one trained with PSSM data only, the neural network using context-based scores as additional features results in improvements in the Q_2 accuracy, which is higher than the reported accuracies, 72-79%, in [7-16].

TABLE II
Comparison of prediction performance of Solvent Accessibility using PSSM only and PSSM with context-based scores on Cull using 7-fold cross validation

	Q_B	Q_E	Q_2
PSSM Only	78.44%	80.61%	79.50%
PSSM+Score	79.21%	82.00%	80.76%

The context-based scores are effective features for the neural network training process. When context-based scores are incorporated, the prediction accuracies are improved on all three benchmarks compared to using PSSM data only.

Figure 4 depicts an example of solvent accessibility prediction on protein 3NRF, chain 'A' listed in CASP9 targets. The first line, underneath the native structure in figure 4, is the amino acid sequence, the second line is the DSSP assignments of each residue, the third line is the predicted solvent accessibility state when using PSSM information for encoding, and the last line is the prediction when incorporating context based scores with PSSM information. An improvement of 6.61% is achieved in this prediction example upon the incorporation of context based scores with PSSM information.

Table II compares the Q2 accuracy between our method and the popularly used solvent accessibility prediction servers including NETASA [8], Sable [9], Netsurf [10], SPINE [6], and ACCpro [11] on benchmarks of CASP9, Manesh215, and Carugo338. To guarantee fairness, we generate a new set of context-based scores by removing all sequences with 25% or higher sequence identity to the sequences in benchmark from Cull16633.

The predictions of the benchmark sets are performed in 2-state for each method. Sable method provides 10-state predictions, with 10% difference among the states of solvent accessibility. Hence the results reported in Table II, for sable method, are using 0.2 and 0.3 thresholds.

We also compare our method with SANN [14] on benchmark of CASP9. The Q_2 accuracy of SANN on CASP9 is 77.86%

such that the Q_B and Q_E are 69.68% and 86.66%, respectively.

We observe that our method outperforms the other methods, where the Q_2 performance is higher in all benchmarks. However, when considering the predictions of the accessibility states individually, SPINE predictions of the buried state (Q_B) outperforms our method, with an average of 1.49% improvement. On the other hand, our method provides a much higher exposed state prediction (Q_E) with an average of 5.75% difference compared to SPINE. Moreover, SANN predictions of the exposed state (Q_E) outperform our method. However, our method outperforms SANN in the buried state (Q_B) and the overall Q_2 predictions.

TABLE III
Comparison of Q2 accuracy between OUR and other popularly used Solvent Accessibility prediction servers including Netsurfp, ACCpro, Sable, SANN, SPINE, and NETASA on benchmarks of CASP9, Manesh215, and Carugo338.

		CASP9	Manesh215	Carugo338
NETASA	**Q_2**	**69.32**	**71.09**	**69.7**
	Q_B	70.86	72.1	72.04
	Q_E	67.59	69.9	67.22
Sable t=0.2	**Q_2**	**78.47**	**79.83**	**78.68**
	Q_B	78.27	80.2	78.48
	Q_E	78.69	79.4	78.91
Sable t=0.3	**Q_2**	**75.13**	**77.04**	**75.94**
	Q_B	89.55	91.08	90.29
	Q_E	59.58	60.35	60.33
Netsurf	**Q_2**	**79.15**	**80.83**	**80.04**
	Q_B	80.04	83.35	81.27
	Q_E	78.19	78.49	78.13
SPINE	**Q_2**	**77.86**	**80.5**	**79.68**
	Q_B	83.22	85.3	85.33
	Q_E	72.08	74.8	73.53
ACCpro	**Q_2**	**76.18**	**78.87**	**77.99**
	Q_B	81.15	83.19	83.12
	Q_E	70.81	73.76	72.41
Our-Method	**Q_2**	**80.82**	**81.93**	**81.14**
	Q_B	81.46	84.27	83.65
	Q_E	80.13	79.14	78.39

IV. CONCLUSIONS

An approach of deriving context-based scores based on the mean-field potentials for characterizing the favorability of residues in solvent accessibility according to their amino acid environment is developed in this paper. Recently, the increasing number of experimentally determined protein structures in PDB has made sufficient number of samples available. This enables us to obtain reliable statistics for mean-field potentials and thus leads to context-based scores with better accuracy. These context-based scores are selected as features together with other sequence and evolutionary information in neural network training for solvent accessibility predictions. The effectiveness of using context-based features has been demonstrated in our computational results in 7-fold cross validation as well as on benchmarks of Manesh215, Carugo338, and CASP9, where enhancements of prediction accuracies are observed.

A web server implementing our solvent accessibility prediction program is currently under development. 2-state with different thresholds, 3-state and 10-state predictions will be added to our method. Also, real value prediction will be part of our future directions in this research.

References

[1] S. Ahmad, M. M. Gromiha, A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," Proteins-Structure Function and Genetics, 50: 629-635, 2003.

[2] W. Kabsch, C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," Biopolymers, 22: 2577-637, 1983.

[3] H. S. Chan, K. A. Dill, "Origins of structure in globular proteins," Proc Natl Acad Sci U S A, 87: 6388-92, 1990.

[4] T. Ooi, M. Oobatake, G. Nemethy, H. A. Scheraga, "Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides," Proc Natl Acad Sci U S A, 84: 3086-90, 1987.

[5] L. Ehrlich, M. Reczko, H. Bohr, R. C. Wade, "Prediction of protein hydration sites from sequence by modular neural networks," Protein Eng, 11: 11-9, 1998.

[6] E. Faraggi, B. Xue, Y. Zhou, "Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network," Proteins, 74: 847-56, 2009.

[7] O. Dor, Y. Q. Zhou, "Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties," Proteins-Structure Function and Bioinformatics, 68: 76-81, 2007.

[8] S. Ahmad, M. M. Gromiha, "NETASA: neural network based prediction of solvent accessibility," Bioinformatics, 18: 819-824, 2002.

[9] R. Adamczak, A. Porollo, J. Meller, "Accurate prediction of solvent accessibility using neural networks-based regression," Proteins-Structure Function and Bioinformatics, 56: 753-767, 2004.
[10] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," Bmc Structural Biology, 9, 2009.
[11] G. Pollastri, P. Baldi, P. Fariselli, R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," Proteins-Structure Function and Genetics, 47: 142-153, 2002.
[12] Z. Yuan, K. Burrage, J. S. Mattick, "Prediction of protein solvent accessibility using support vector machines," Proteins-Structure Function and Genetics, 48: 566-570, 2002.
[13] H. Kim, H. Park, "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor," Proteins, 54: 557-62, 2004.
[14] K. Joo, S. J. Lee, J. Lee, "Sann: Solvent accessibility prediction of proteins by nearest neighbor method," Proteins-Structure Function and Bioinformatics, 80: 1791-1797, 2012.
[15] J. Sim, S. Y. Kim, J. Lee, "Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method," Bioinformatics, 21: 2844-2849, 2005.
[16] H. Naderi-Manesh, M. Sadeghi, S. Arab, A. A. M. Movahedi, "Prediction of protein surface accessibility with information theory," Proteins-Structure Function and Genetics, 42: 452-459, 2001.
[17] M. J. Thompson, R. A. Goldstein, "Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes," Proteins-Structure Function and Genetics, 25: 38-47, 1996.
[18] B. Rost, C. Sander, "Conservation and Prediction of Solvent Accessibility in Protein Families," Proteins-Structure Function and Genetics, 20: 216-226, 1994.
[19] J. A. Cuff, G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," Proteins-Structure Function and Genetics, 40: 502-511, 2000.
[20] J.-K. H. W.-L. Hsu. (2006 SAS prediction server. Available: http://140.113.239.214/~weilun/index.php
[21] M. J. Sippl, "Calculation of Conformational Ensembles from Potentials of Mean Force - an Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins," Journal of Molecular Biology, 213: 859-883, 1990.
[22] J. L. Sussman, D. W. Lin, J. S. Jiang, N. O. Manning, J. Prilusky, O. Ritter, E. E. Abola, "Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules," Acta Crystallographica Section D-Biological Crystallography, 54: 1078-1084, 1998.
[23] O. Carugo, "Predicting residue solvent accessibility from protein sequence by considering the sequence environment," Protein Engineering, 13: 607-609, 2000.
[24] L. Kinch, S. Y. Shi, Q. Cong, H. Cheng, Y. X. Liao, N. V. Grishin, "CASP9 assessment of free modeling target predictions," Proteins-Structure Function and Bioinformatics, 79: 59-73, 2011.
[25] G. L. Wang, R. L. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19: 1589-1591, 2003.
[26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, 25: 3389-402, 1997.
[27] F. Vonderviszt, G. Matrai, I. Simon, "Characteristic Sequential Residue Environment of Amino-Acids in Proteins," International Journal of Peptide and Protein Research, 27: 483-492, 1986.