Prediction of Protein B-Factors Using Multi-Class Bounded SVM

Peng Chen[1,2,3], Bing Wang[1,2,3], Hau-San Wong[3] and De-Shuang Huang[1,*]

[1]*Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei Anhui, 230031, China;* [2]*Department of Automation, University of Science and Technology of China, Hefei Anhui, 230026, China;* [3]*Department of Computer Science, Virtual Reality Laboratory, Image Computing Group, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong*

Abstract: In this paper, we propose the adoption of the bounded support vector machine (BSVM) to predict the B-factors of residues based on a number of distinctive properties of residues. Due to the ability of multi-class classification of the BSVM, we can elaborately distinguish our targets and obtain relatively higher accuracy.

Keywords: B-factors, Bounded Support Vector Machine, Sequence profile, Evolutionary rate, Hydrophobicity profile.

1. INTRODUCTION

Despite significant improvements of structure determination techniques, the gap between the number of proteins for which structure is deposited in public databases, e.g., PDB [1], and the number of proteins for which sequences are known, is increasing more and more rapidly. Methods that predict aspects of protein structure continue to be the crucial means of obtaining structural information that helps to unravel the protein function [2-3]. Since it is difficult to directly predict the 3D protein structure, we propose to study specific protein structural behavior, such as protein flexibility, in this work. In general, protein flexibility is an important determinant of protein structural behavior, and experimental evidence shows that the effect of protein flexibility is predominant both in solution and in the solid state. The temperature factor (or B-factor), as determined using crystallography, is linearly related to the mean square displacement of an atom and indicates the atomic flexibility in the crystalline state. Previous works have shown that hydrophobic residues, which are usually buried, tend to be more rigid whereas charged residues tend to be more flexible [9-10]. Moreover, it is well known that long loops tend to be more flexible than regular secondary structures such as helices and strands. Another observation indicates that residues in regular secondary structure (helix and strand) tend to occupy regions of lower normalized B-factors, while residues in nonregular secondary structure occupy those with higher normalized B-factors [25]. More importantly, flexibility provides important information about protein structure and function. Halle showed that the packing density, or contact density, which describes the number of noncovalent neighbor atoms within a local region of *~1.5 nm*^{3} volume, is inversely proportional to atomic thermal motion or B-factor [26]. These earlier discussions indicate that the prediction of B-factors can help to unravel protein function and further predict protein structure [9, 10, 25, 26].

Researchers have proposed a number of different approaches for predicting the B-factors of protein crystal structures. For example, Bahar [4] proposed a Gaussian network model (GNM) to address this problem. It is well known that B-factor represents the extent of free movement of atom, and that higher atomic B-factor correlates with their exposed surface area [5]. As a result, Haliloglu [6] refined the GNM model and analyzed the vibrational motion of globular proteins. Kundu *et al.*[7] found that the GNM using C_α atomic coordinates in an isolated molecule could achieve a correlation coefficient (CC) of 0.59 between predicted and experimental B-factors. Most of other previous works were focused on the so-called “classification problem.” The classification of B-factors was to assign the residues to one of two states, i.e., rigid or flexible, with arbitrary B-factors cutoff thresholds [8, 9]. Specifically, the B-factors are classified into two states, i.e., low and high B-factors, and then four regions are analyzed, i.e., low-B-factor ordered regions, high-B-factor ordered regions, short disordered regions, and long disordered regions [10]. The threshold for these two classes is mostly determined from experience. Currently, a good B-factor prediction approach is based on support vector regression (SVR) proposed by Yuan *et al.*[11]. Although the SVR approach can achieve a prediction accuracy of 70% and a CC of 0.53, it can only classify the residues for a given protein into two classes, i.e., rigid or flexible. At the same time, a higher variance of CCs is observed across different proteins. Another drawback of the SVR approach is that it does not capture the global information of the amino acid sequence due to the limited size of the local neighborhood [12].

The bounded support vector machine (BSVM) approach in this work is somewhat related to the SVR classifier, but can achieve better B-factor prediction for each residue due to its capability to solve multi-class problems. This work is based on the observation that our classifier took in this article can classify three or more class states. Due to the more class states classification, we can elaborately distinguish our targets and obtain relatively higher accuracy. Unlike previous works, we have also utilized the observation that multiple classifier systems gain widespread attention due to its ability to improve classification rate by fusing the outputs of

[*]Address correspondence to this author at the Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei Anhui, 230031, China; E-mail: dshuang@iim.ac.cn

0929-8665/07 $50.00+.00

a number of independent classifiers. [27, 28] As a result, our objective is to solve the problem of how these three B-factor states are to be correctly classified, and to obtain the predicted B-factors based on BSVM.

In this paper, we first encode the three B-factor classes for the BSVM classifier in terms of three characteristics of each residue: sequence profile, evolutionary rate, and hydrophobicity profile. Second, three predictors are constructed which are respectively called as the 'SP predictor', which performs prediction based on the first characteristic of each residue, the 'SE predictor', which performs prediction based on the first two characteristics, and the 'SEH predictor', which involves all the above characteristics. Finally, we apply the greedy kernel principal component analysis (GKPCA) approach [13] to reduce the complexity of the training data as well as the computation and memory requirements of the BSVM algorithm. The comparative results among these three predictors show that the SEH predictor obtains the highest performance, and that it also surpasses the other approaches for predicting B-factors, such as SVR, PROFboval [25], and other techniques. In particular, the experimental results show that the average CC of the SEH predictor is up to 0.60, while the average accuracy of classification also surpasses 59% for this three class problem.

2. MATERIALS AND METHODS

2.1. Materials and Datasets

We obtained 776 protein chains using PDB-REPRDB [14], which is a database of protein chains from PDB based on PDB Rel. #2005_05_29, and updated on 10th June 2005. We selected those chains from different proteins which are resolved by X-ray crystallography with resolution *≤2.0Å*, R-factor *≤19%*. The sequence identity between two selected chains is less than 25%. As a result, we have two chain groups: one contains chains whose proteins have only one chain, and another contains chains whose proteins consist of more than one chain. The former group has 352 chains, while the latter has 424 chains. Subsequently, a further refinement step is performed. For the single–chain group we retain 287 chains, where three chains are excluded since their C_α B-factors are either equal to zero or a constant, and 62 chains without Consurf-Hssp [17] files are also removed. In the same way, we select 332 chains for the multi-chain group. As a result, our two chain groups consist of 619 (*287+332*) chains.

To validate our approach, we apply 2-fold cross-validation tests in our experiments for each chain group. Specifically, we split each chain group into a training set and a test set with approximately the same number of chains. Each set is used in turn as training and test set. The training and testing of each BSVM was carried out twice using one set for training and the other set for testing.

2.2. Methods

To predict the B-factors of each protein chain, the B-factor of C_α atom may be chosen to represent that of each amino acid. The B-factors can be classified into three states (or classes), i.e., high-B-factor, median-B-factor, and low-B-factor. As a result, two threshold values, t_{h-m} and t_{m-l}, are required to separate these three classes. In general, the input for the BSVM to predict B-factors is a set of training vectors, $\Upsilon_{XY} = \{(x_1, y_1), ..., (x_c, y_c)\}$, where $x_i \in \chi \subset \mathbb{R}^n$, and the corresponding output takes value in the set $y \in \Upsilon = \{1,2,3\}$.

In terms of the method described in [15] and its subsequent refinement [16], a dual formulation of the BSVM can be derived. The parameters of the multi-class rule can be found by solving the multi-class BSVM problem based on the following expression:

$$A^* = \arg\max_{A} \sum_{i\in I} \sum_{y\in\Upsilon\backslash\{y_i\}} \alpha_i^y - \frac{1}{2}\sum_{i\in I} \sum_{y\in\Upsilon\backslash\{y_i\}} \sum_{j\in I} \sum_{u\in\Upsilon\backslash\{y_j\}} \alpha_i^y \alpha_j^u h(y,u,i,j) \quad (1)$$

subject to

$$\alpha_i^y \geq 0, \qquad i \in I, y \in \Upsilon \backslash \{y_i\}$$

where $A = [\alpha^1, \alpha^2, \alpha^3]$ is the optimized weight vector set, and the function $h : \Upsilon \times \Upsilon \times I \times I \rightarrow \mathbb{R}$ is defined as:

$$h(y,u,i,j) = (< x_i \bullet x_j > +1)(\delta(y_i, y_j) + \delta(y,u) - \delta(y_i,u) - \delta(y_i,y)) + \frac{\delta(y,u)\delta(i,j)}{2C} \quad (2)$$

The multi-class classifier is composed of the set of discriminant functions $f_y : \chi \rightarrow \mathbb{R}$ which are evaluated as:

$$f_y(x) = \sum_{i\in I} < x_i \bullet x > \sum_{u\in\Upsilon\backslash\{y_i\}} \alpha_i^u (\delta(u,y_i) - \delta(u,y)) + b_y, \qquad y \in \Upsilon \quad (3)$$

where the bias $b_y, y \in \Upsilon$ is given by:

$$b_y = \sum_{i\in I} \sum_{u\in\Upsilon\backslash\{y_i\}} \alpha_i^u (\delta(y,y_i) - \delta(y,u)), \qquad y \in \Upsilon$$

The non-linear (kernel) classifier can be obtained by substituting the selected kernel $k : \chi \times \chi \rightarrow \mathbb{R}$ for the inner products $< x \cdot x' >$ to Eqs. (2) and (3). The kernel function is usually chosen to be the radial basis function (RBF). Finally, we can perform classification based on the multi-class classification rule $q : \chi \rightarrow \Upsilon = \{1,2,3\}$, where:

$$q(x) = \arg\max f_y(x), \qquad y \in \Upsilon \quad (4)$$

2.3. Encoding Scheme for BSVM

We first encode the input vectors of the BSVM classifier for each residue site based on a sliding window of residues centered at the current site except near the N- and C-termini. Each sliding window consists of a continuous set of residues with an odd size of *win*, where *win* is set to 13 in our work.

For the input of the SP predictor, we adopt the residue sequence profile obtained from the HSSP database [24]. Each residue is represented by 20 attributes whose values are determined from multiple sequence alignment and their potential structural homologs. Therefore, a training vector for one residue site contains *20*13=260* elements, as *win* is equal to 13.

For the SE evolutionary rate based predictor, the evolutionary rate is characterized by taking into account the phylogenetic relationships between the homologs and the stochastic nature of the evolutionary process, so that the conservation level for each residue can be inferred by using the maximum likelihood (ML) criterion [17]. Each evolutionary rate score can then be appended to the first 20 sequence profile elements for each residue. As a result, the input vector of SE predictor therefore contains *21*13=273* elements which contain elements from sequence profile and evolutionary rate.

For the SEH predictor, each hydrophobicity value can be appended to the first 21 elements of training vector for each residue. In general, hydrophobicity plays a very important role in characterizing the physico-chemical property in different areas of chemistry, medicine, and pharmacology [18, 19]. The input vector thereby contains *22*13=286* elements totally.

2.4. Normalization Scheme

It is necessary for us to normalize the B-factors to equalize its range. The method used in this article was derived from Karplus *et al.*[20]. The normalized data y' can be obtained by the following equation:

$$y' = \frac{y-\mu}{\sigma} \tag{5}$$

where μ and σ denote the mean and the standard deviation of the original data, *y,* respectively.

2.5. Evaluation Measures for Performance of Predictors

To verify our method, we calculated the Pearson correlation coefficient (CC) between the theoretical and experimental B-factors by the following equation:

$$CC = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{\sum(x-\mu_x)^2 \sum(y-\mu_y)^2}} \tag{6}$$

where *x* and *y* are the theoretical and experimental B-factors of C_α atom, respectively. *E* is the mathematical expectation and $\mu_x = Ex, \mu_y = Ey$.

In addition, other measurements have been introduced to evaluate the performance of the predictors. Here, we applied the criteria of accuracy (Acc) and coverage (Cov) which are adopted at CASP/CAFASP [21, 22]. These are defined as follows:

$$Acc = \frac{TP}{TP+FP} \quad Cov = \frac{TP}{TP+FN} \tag{7}$$

where *TP* denotes the number of true positives, i.e., the B-factors that are assigned to the correct classes to which they actually belong; *FP* denotes the number of false positives, i.e., the B-factors that are assigned to another class which in fact belong to the current class; and FN stands for the number of false negative, i.e., the B-factors that are assigned to the current class which in fact belong to another class.

3. RESULTS AND DISCUSSIONS

To reduce the complexity of the BSVM algorithm, greedy kernel principal component analysis (GKPCA) is applied to preprocess the training data of the BSVM algorithm. This approach aims at representing data in a low dimensional space with possibly minimal representation error similar to the Principal Component Analysis (PCA). The basis vectors of the GKPCA for data representation are selected from the training set. These basis vectors can be selected by a simple greedy algorithm which does not require eigenvalue decomposition (as the PCA does) and is of complexity $O(nm^3)$, where *n* is the size of the training set and *m* is the number of the basis vectors. It can be shown that GKPCA can significantly reduce the number of support vectors while maintaining the accuracy of the classifiers. As a result, by applying the GKPCA algorithm, we can obtain a set of 36-dimensional training data for the SP predictor, 49-dimensional (*36+13*) data for SE predictor, and 62-dimensional (*36+13+13*) data for SEH predictor.

When evaluating Eq. (3), the BSVM algorithm obtains two bounds, the upper bound F_{UB} and the lower bound F_{LB} on the optimal value, $F(\mathbf{A^*}, \boldsymbol{b}^*)$. As a result the bounds can be applied to define the following two stopping conditions, i.e., the relative tolerance, $\varepsilon_{rel} \geq (F_{UB} - F_{LB})/(F_{LB}+1)$ and the absolute tolerance, $\varepsilon_{abs} \geq F_{UB}$. In our article, when the two stopping conditions based on $\varepsilon_{rel} = 0.001$ and $\varepsilon_{abs} = 0.0$ are satisfied, the algorithm will be terminated.

The predicted accuracies of the BSVM approach on the single-chain group and multi-chain group are shown in Fig. **1**. To validate our results, we apply 2-fold cross-validation tests for each group. The comparison between our proposed approach, the SVR [11], and the PROFboval method is shown in Tables **1** and **2**.

From the left graph of (Fig. **1**), it is apparent that the predictors trained on the single-chain group perform better than those on multi-chain group. In particular, the SEH predictor is able to attain an accuracy of nearly 91% and a CC of 78%, which is approximately equal to the prediction limit CC of 81% calculated by Radivojac [10], on the training set. The average accuracy and CC for the multi-chain group are 82% and 69.5%, respectively.

The right graph of (Fig. **1**) gives a comparison based on the test set. Similar conclusions to that on the training set may be drawn. In summary, the SEH predictor yields an accuracy of 65.7% and a CC of 58.7% on the single-chain group. For the multi-chain group, the SEH predictor also obtains an accuracy of 52.7% and CC of 54.8%. From the experimental results, it is also seen that the SE predictor can obtain higher accuracies than the SP predictor. As a result, the SEH predictor achieves the best performance among the three predictors.

Table 1. The Comparison of Average Accuracies and CCs between SVR and Our Three Predictors Based on the Chosen Chains

Sequence Length	SP	SE	SEH	SVR
L≤150(162/226)[a]	0.59(0.53)	0.582(0.554)	0.63(0.60)	(0.533)
150<L≤300(249 /303)	0.575(0.515)	0.545(0.53)	0.585(0.57)	(0.534)
L>300(208/237)	0.535(0.488)	0.525(0.51)	0.56(0.532)	(0.523)
All(619/766)	0.567(0.52)	0.551(0.531)	0.592(0.568)	0.70(0.53)[b]
2-states comparison[c]	0.68(0.52)	0.713(0.531)	0.741(0.568)	0.70(0.53)

[a] The bracketed numbers denote the number of protein peptide chains in our data set and the SVR set respectively.
[b] Due to a lack of detailed information about the average accuracies of the SVR approach, the corresponding accuracies with respect to sequence length are not included in this table.
[c] There, we convert our 3-states prediction to a 2-states prediction that can directly compare our approach to the SVR method.

Table 2. The Comparison of Average Accuracy, Coverage and CC between the PROFboval and Our Three Predictors Based on the Chosen Chains

Prediction method	Acc	Cov	CC
PROFboval	70.1±0.18	73.9±0.17	0.4[a](0.50) [b]
SP predictor	56.7±0.24	61±0.21[d]	0.52[c]
SE predictor	55.1±0.18	63.4±0.20	0.531
SEH predictor	59.2±0.21	65±0.20	0.568
2-states comparison[e]	0.741±0.23	78±0.21	0.568

[a] The number denotes the Pearson correlation coefficient calculated by PROFboval on their own dataset.
[b] The bracketed number denotes the correlation coefficient calculated by PROFboval on the dataset used in SVR method [11].
[c] This CCs stands for the correlation coefficient of our method on our integrated dataset.
[d] The ± values mark the standard errors compiled over our dataset.
[e] There, we convert our 3-states prediction for SHE predictor to a 2-states prediction that can directly compare our approach to the SVR method.

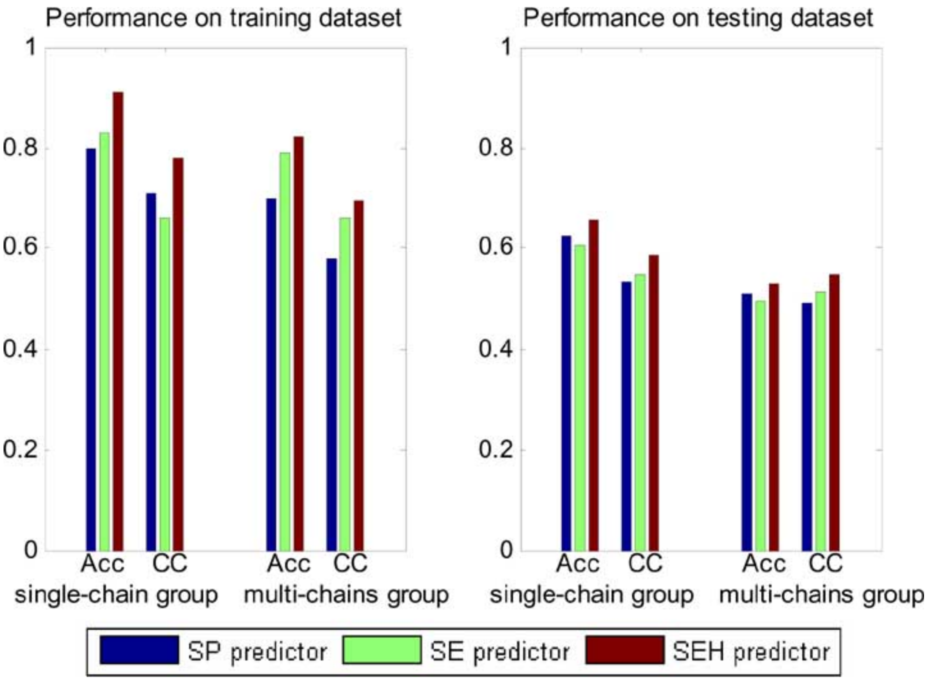

Figure 1. Comparison of predictor performance on the training and test sets. The first grey bar denotes the SP predictor performance, the second bar the SE predictor performance, and the third black bar the SEH predictor performance. Acc stand for the accuracy of each predictor, while CC stands for the Pearson correlation coefficient as defined in Eq. (5). The left graph shows the performance on training set, while the right graph shows the performance on test set.

In general, the adoption of the multi-chain group as training data will lead to a worsening of the predictor performances (in terms of accuracies and CCs) by around 4% compared to the single-chain group. Further investigation can provide a clue to why the performance becomes worse for the multi-chain group: If we consider the spatial structure of two peptide chains in a multi-chain protein, they may interact with each other due to their proximities. This interaction may restrain the thermal motion of the residues close to this chain-chain interface. Determining how to improve the prediction accuracy of the B-factors of residues which are located at or near to the chain-chain interface in a multi-chain protein structure will be our future research problem.

Compared to the SVR approach, we can calculate the accuracies and the CCs in terms of sequence length by integrating the single-chain and multi-chain groups. The comparison between the performance of the SVR approach and our three predictors is shown in Table **1**.

Due to the increased complexity of performing 3-state prediction of B-factors in our BSVM approach, the accuracy is less than the 2-state prediction result based on SVR. However, our method achieves higher CCs than the SVR method, due to the ability of the BSVM method to capture more global information of the amino acid sequence. For the 3-states of B-factors we obtain a CC of 0.568, while better performance is observed in the shorter sequences whose lengths are less than 150 for the SEH predictor, with a corresponding CC of 0.602. When converting our 3-states prediction to a 2-states prediction, our predictor performs higher accuracy of 0.741 than SVR method with a accuracy of 0.70. When the training data of B-factors are classified into more than three states, it can improve the prediction CCs slightly, but the corresponding complexity and computation of this multi-state problem can increase rapidly.

Observation based on the PROFboval [25] method published by Rost and co-workers shows that the best predictor can reach an accuracy of 70.1% and a coverage rate of 73.9%. Because of the difference in the data sets used in our method and the PROFboval method, only an indicative indirect comparison of prediction performance can be provided in Table **2**. Similar to the comparison in (Table **1**), due to the complexity of 3-state prediction, accuracies and coverage rates of our 3 classes BSVM method are less than those of the PROFboval method, but a higher Pearson correlation coefficient is again obtained due to our more sophisticated prediction approach. As converting our 3-states prediction to a 2-states prediction, our classifier performs higher accuracy of 0.741 and coverage of 0.78 than the PROFboval method which only obtains accuracy of 0.701 and coverage of 0.739.

In order to demonstrate the meanings of different CC values, (Fig. **2**) illustrates the theoretical and the experimental B-factors for two protein chains, the PDB entry 1bxa and chain 'A' of the PDB entry 1cku, which belong to the single-chain and multi-chain group respectively. PDB:1bxa denotes a protein of amicyanin that consists of one chain with 105 residues; whereas the PDB: 1cku_A is the protein chain 'A' with 85 residues, selected from the High-Potential Iron Protein (HIPIP) which consists of two chains, 'A' and 'B'. From (Fig. **2**), it can be seen that the accuracies of these proteins can be up to 73.2% and 74%, and the CCs of 0.824 and 0.856 for protein PDB: 1bxa and protein chain PDB: 1cku_A, respectively.

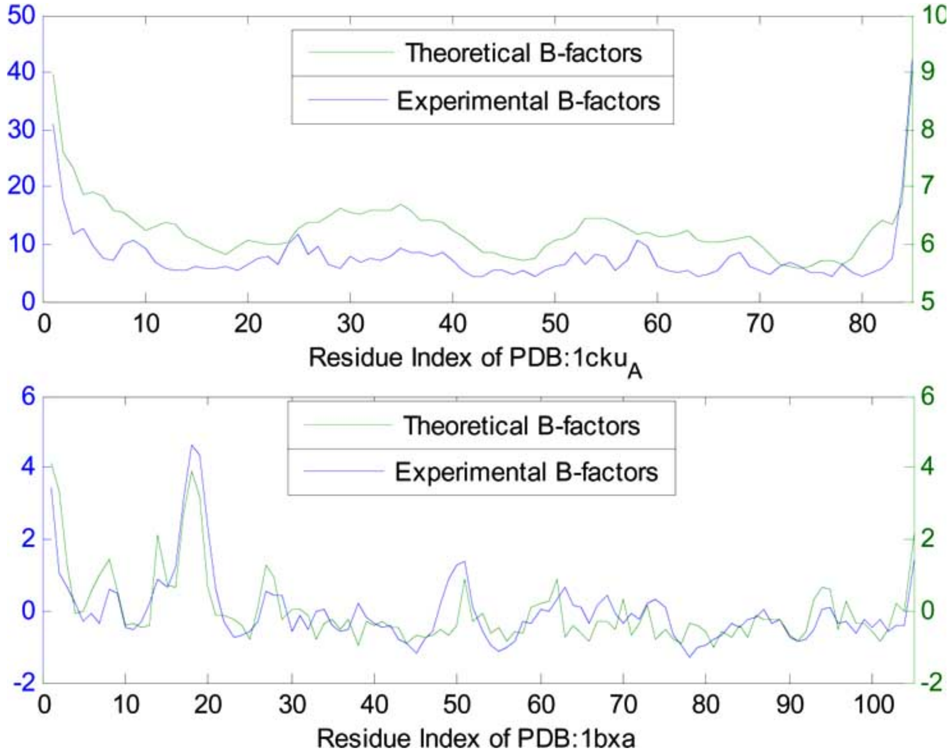

Figure 2. The comparison of theoretical (grey) and experimental (black) B-factors for protein PDB: 1bxa and a chain 'A' of protein PDB:1cku. The sequence length of protein PDB:1cku and 'A' chain of protein PDB:1bxa are 85 and 105, respectively.

Because the protein PDB:1cku consists of two peptide chains, 'A' and 'B', the residues ALA16 and THR 13 in the 'A' chain interact with their spatial neighboring residues ALA12 and ALA16 in 'B' chain. Due to these interactions, their flexibilities can be impaired to a great extent. Generally, in order to predict the B-factors we usually consider a single peptide chain where no chain-chain interaction is involved which may affect the prediction accuracy. This type of interaction is illustrated in (Fig. **3**).

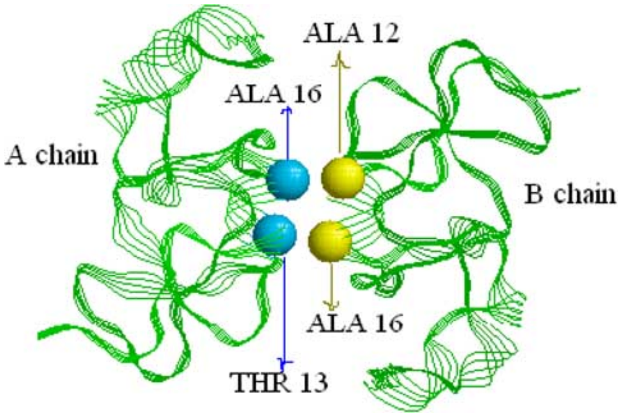

Figure 3. The chain-chain interaction between two chains of protein PDB:1cku. Interactions between the residues ALA16 in 'A' chain and ALA12 in 'B' chain, as well as the residues THR13 in 'A' chain and ALA16 in 'B' chain, might impair the flexibility of each residue.

To perform classification, the two thresholds of the 3-states of B-factors are set to 25% and 78%. By varying the thresholds, the prediction accuracies may change by 3.4% and the CCs by 5%. Another parameter for the BSVM, the regularization constant *C*, is set to 5.0 in our experiments, and its effect is described in [23]. In our approach, the stopping conditions are defined based on a relative tolerance of 0.001 and an absolute tolerance of zero. In general, increasing the relative tolerance may lead to a decrease in prediction accuracy, while decreasing the relative tolerance may improve the performance with an accompanying increase in computation time.

CONCLUSION

In this paper, we propose the adoption of the multi-class BSVM approach for predicting B-factors from protein sequence based on sequence profile, evolutionary conservation, and hydrophobicity profile. The experimental results show that the best prediction is obtained in the case of the SEH predictor with corresponding CC values of 0.587 and 0.548 for the single-chain and multi-chain group respectively, and the average CC is 0.568 which outperforms the SVR approach with a corresponding average CC of 0.53. For future works, we shall improve the classification accuracy, and consider how to apply the B-factor prediction results to structural prediction, such as contact map prediction.

Received: November 08, 2006 Revised: December 11, 2006 Accepted: December 12, 2006

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (Nos. 60472111, and 30570368), the 863 Project of National High Technology of China (2006AA 02Z309), and a grant from the Research Grants Council of Hong Kong Special Administrative Region, China (No. CityU 1160/04E).

REFERENCES

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., T., Bhat, T. N., Weissig, H., Shindyalov, I.N., Bourne, P. E. (**2000**) *Nucl. Acids Res.*, *28*, 235-242.

[2] Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J.S., Skolnick, J. and Godzik, A. (**2000**) *Protein Sci.*, *8*, 1104-15.

[3] Thornton, J. M. (**2001**) *Science*, *292*, 2095-2097.

[4] Bahar, I., Atilgan, A. R. and Erman, B. (**1997**) *Folding Des.*, 2, 173.

[5] Carson, M., Buckner, T.W., Yang, Z., Narayana, S.V.L. and Bugg, C.E. (**1994**) *Acta Cryst.*, D50, 900-909.

[6] Haliloglu, T., Bahar, I. and Erman, B. (**1997**) *Phys. Rev. Lett.*, *79*, 3090.

[7] Kundu, S., Melton, J.S., Sorensen, D.C. and Phillips, G.N. (**2002**) *Biophys J.*, *83*, 723–732.

[8] Karplus, P.A. and Schulz, G.E. (**1985**) *Naturwissenschaften*, *72*, 212-213.

[9] Vihinen, M., Torkkila, E. and Riikonen, P. (**1994**) *Proteins*, *19*, 141–149.

[10] Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., Dunker, A.K. (**2004**) *Protein Sci.*, *13*, 71–80.

[11] Yuan, Z., Bailey, T. L. and Teasdale, R. D. **(2005)** *PROTEINS*, *58*, 905-712

[12] Nguyen, M.M. and Rajapakse, J.C. (**2003**) *Genome Informatics*, *14, 218–227.*

[13] Franc, V. and Hlavac, V. (**2003**) In N. Petkov and M.A. Westenberg, editors, *Computer Analysis of Images and Patterns*, 426–433, Berlin, Germany, Springer.

[14] Noguchi, T. and Akiyama, Y. (**2003**) *Nucleic Acids Res.*, 31, 492-493.

[15] Hsu, C.W. and Lin, C.J. (**2002**) *IEEE Trans. Neural Networks*, *13*, March 2002.

[16] Franc, V. and Hlavac, V. (**2002**) In: R. Kasturi,D. Laurendeau, and Suen C., editors, *16th International Conference on Pattern Recognition*, *2*, 236–239.

[17] Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. and Ben-Tal, N. (**2005**) *PROTEINS: Structure, Function, and Bioinformatics, 58, 610-617.*

[18] Lyman, W. J., Reehl, W. F. and Rosenblatt, D. H. (**1982**) *Handbook of Chemical Property Estimation Methods.*

[19] Kyte, J. and Doolittle, R.F. (**1982**) *J. Mol. Biol.*, *157*, 105-132.

[20] Karplus, P.A. and Schulz, G.E. (**1985**) *Naturwissenschaften*, *72*, 212–213.

[21] Fischer, D., Rychlewski, L., Dunbrack, R. L. Jr., Ortiz, A. R. and Elofsson, A. (**2003**) *Proteins: Structure,Function, and Genetics*, *53*, 503-516.

[22] Punta, M. and Rost, B. (**2005**) *Bioinformatics*, *21*, 2960 - 2968.

[23] Franc, V. (**2005**) PhD Thesis, Center for Machine Perception, *K13133 FEE Czech Technical University*, Prague.

[24] Dodge, C., Schneider, R. and Sander, C. (**1998**) *Nucleic Acids Res.*, *26*, 313–315.

[25] Schlessinger, A. and Rost, B. (**2005**) *PROTEINS: Structure, Function, and Bioinformatics*, *61, 115-126*.

[26] Halle, B. (**2002**) *Proc Natl Acad Sci USA*, *99*, 1274-1279.

[27] Kittler, J. and Alkoot, F.M. (**2003**) *IEEE Trans. Pattern Anal. Mach. Intell*, *25*, 110–115.

[28] Kuncheva, L.I. **(2004)** Combining pattern classifiers: methods and algorithms. U.S.: Wiley.