Prediction of Protein Flexibility Profile

Article

3 authors, including:

Zheng Yuan
The University of Queensland
25 PUBLICATIONS **1,054** CITATIONS

SEE PROFILE

Rohan D Teasdale
The University of Queensland
178 PUBLICATIONS **10,093** CITATIONS

SEE PROFILE

Prediction of Protein Flexibility Profile

Zheng Yuan*, Timothy L. Bailey, and Rohan D. Teasdale

Institute for Molecular Bioscience and ARC Centre in Bioinformatics,

The University of Queensland, St. Lucia, 4072, AUSTRALIA

*Corresponding author

Address for correspondence:

Institute for Molecular Bioscience,

The University of Queensland,

St. Lucia, 4072, AUSTRALIA

Phone: +61 7 3346 2633, FAX: +61 7 3346 2101

Email: z.yuan@imb.uq.edu.au

Abstract

The polypeptide backbones and the side chains of proteins are constantly moving due to thermal motion or kinetic energy of the atoms. The B-factors of protein crystal structures reflect the fluctuation of atoms about their average positions and provide important information on protein dynamics. Computational approaches to predict the thermal motion are useful for analysing the dynamic properties of proteins with unknown structures. In this paper, we utilize a novel support vector regression (SVR) approach to predict the B-factor distribution (flexibility profile) of a protein from its sequence. We explore schemes for encoding sequences and various settings for the parameters used by SVR. Based on a large dataset of high-resolution proteins, our method predicts the B-factor distribution with a Pearson correlation coefficient of 0.53. In addition, our method predicts the flexibility profile with a correlation coefficient of at least 0.56 for more than half the proteins. Our method also performs well for classifying residues (rigid vs. flexible). For almost all choices of predicted B-factor threshold, prediction accuracies (percent correctly predicted residues) are greater than 70%. These results exceed the best results of other sequence-based prediction methods.

Keywords

B-factor; flexibility; support vector regression; evolutionary information; ROC analysis; protein sequence

Introduction

Protein structures are not static and rigid. The polypeptide backbones, and especially the side chains, are constantly moving due to thermal motion or kinetic energy of the atoms (Brownian motion). Protein internal motion or flexibility is in high correlation with protein functions such as catalysis and allostery.[1] Thermal motion (displacement of functional groups) is one of the most important motions within protein structures.

The B-factor (atomic displacement parameter) in protein crystal structures reflects the fluctuation of an atom about its average position. The distribution of B-factors along a protein sequence is regarded as an important indicator of the protein's structure, reflecting its flexibility and dynamics. A large B-factor indicates high mobility of individual atoms and side chains. B-factors provide a solid experimental source of information on the dynamics of a protein. B-factors have been applied to a variety of problems such as predicting protein flexibility,[2,3] studying protein thermal stability,[4,5] analysing active sites,[6-8] correlating side-chain mobility with conformation,[9,10] analysing protein disordered regions[11,12] and investigating protein dynamics.[13] B-factors have been predicted from protein sequences,[2,3,12] protein atomic coordinates[14] and protein electron density maps.[15]

Measured B-factors in different known structures may be on different scales owing to the applications of different refinement procedures.[16] Instead of raw data, normalized data is usually used to compare B-factors of different protein chains and structures.[2,3,17,18]

Since high-throughput experiments have rapidly generated a large number of protein sequences, fast and accurate computational methods to annotate their structural, functional and dynamic properties are highly desired. Numerous prediction methods have been developed during the last few decades to predict secondary

structure, solvent accessibility and subcellular localization. In contrast, few prediction methods for protein flexibility or mobility exist.

In this study, we put forward a new method to predict B-factor distributions based on a support vector regression (SVR) [19-21] approach. This approach is often effective when the input data is of high dimension and the function to be predicted is highly non-linear. SVR has been successfully applied to some biological problems such as simulating the age of *Drosophila* embryo,[22] analysing microarray data[23] and predicting protein solvent accessible areas.[24]

Most previous work in prediction of B-factors has been on the so-called "classification problem." The classification problem assigns residues to one of two states—rigid or flexible—with arbitrary B-factor cutoff thresholds.[2,3,12] The selection of thresholds is neither objective nor optimal, and the decomposition of residues into two classes decreases the prediction accuracy. To overcome these disadvantages, we predict the B-factor for each residue. That is, our method predicts a series of real values representing a protein sequence (also regarded as the flexibility profile).

Our new method has been examined on a well-prepared non-redundant dataset by cross-validation tests. Our method achieves a Pearson's correlation coefficient (CC) of 0.53, which exceeds all previously reported results. In the two-class problem, our method achieves classification accuracy greater than 70% regardless of the B-factor threshold used to define the two classes.

Methods

Support vector regression (SVR)

The objective of the regression problem is to estimate an unknown continuous-valued function, $y = \hat{f}(X)$, based on a finite number of samples. In our problem, the training samples are represented by the paired values $\{(X_i, y_i)\}$ $(i = 1,\ldots,M)$, where the feature vector X_i characterizes a residue in a protein sequence and y_i represents its associated normalized B-factor value. We estimate the true function, $\hat{f}(X)$, by the following function using support vector regression to find optimum weights, *W*, and bias, *b*,

$$f(X_i) = \langle W, \Phi(X_i) \rangle + b. \tag{1}$$

Since $\Phi(X_i)$ may be a non-linear mapping from the input space to a (possibly higher dimensional) feature space, SVR is able to perform non-linear regression.

The regression parameters (*W* and *b*) are optimised to minimize an objective function that combines the norm of the weights, $\|W\|^2$, and an empirical risk function. The risk function is based on Vapnik's ε-insensitive loss function,

$$L_\varepsilon(y_i - f(X_i)) = \begin{cases} 0 & \text{if} \quad |y_i - f(X_i)| \le \varepsilon \\ |y_i - f(X_i)| - \varepsilon, & \text{otherwise} \end{cases}. \tag{2}$$

Only those errors larger than ε contribute to the loss function. The risk function is defined as

$$R_{emp}(W,b) = \frac{1}{M}\sum_{i=1}^{M} L_\varepsilon(y_i - f(X_i)). \tag{3}$$

Slack variables ξ and ξ^* are introduced to measure the deviation of samples outside the ε-insensitive zone. Using the slack variables, the objective function for SVR is defined as

$$F(W,b) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*), \tag{4}$$

where C is a regularization constant that determines the trade-off between training errors and model complexity.

With the foregoing definitions in place, we can formulate the SVR problem as the following constrained optimisation problem:

$$\text{minimize} \quad F(W,b)$$

$$\text{subject to} \begin{cases} f(X_i) - y_i \le \varepsilon + \xi_i \\ y_i - f(X_i) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \ \text{ for } i = 1,\ldots,M \end{cases} \tag{5}$$

To solve the optimisation problem, Lagrange multipliers α_i and α_i^* are added to the condition equations and the above problem can be written as its dual form:

maximize

$$-\varepsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)y_i - \frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\left\langle \Phi(X_i), \Phi(X_j) \right\rangle$$

$$\text{subject to } 0 \le \alpha_i, \alpha_i^* \le C \text{ and } \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0 \tag{6}$$

Only the non-zero values of Lagrange multipliers are useful in predicting the regression line and their corresponding samples are known as support vectors.

Notice that in this form, the solution only requires that the dot product of the transformed input vectors be known. So the user of SVR does not choose the mapping function $\Phi(X)$ explicitly. Instead, the user chooses a so-called kernel function whose value computed on two input points equals the dot product of the transformed points in the feature space. A suitable function,

$$K(X_i, X_j) = \left\langle \Phi(X_i), \Phi(X_j) \right\rangle, \tag{7}$$

can be found under general conditions. Choices for the kernel function include polynomial functions,

$$K(X_i, X_j) = (\langle X_i, X_j \rangle + 1)^n, \tag{8}$$

And, radial basis functions (RBF),

$$K(X_i, X_j) = \exp(-\gamma \| X_i - X_j \|^2). \tag{9}$$

The regression function in Eq. (1) can be rewritten in terms of the kernel function, the Lagrange multipliers and the bias term as

$$f(X) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) K(X_i, X) + b. \tag{10}$$

Once the Lagrange multipliers and the bias, *b*, are determined from the training samples, Eq. (10) can be used to predict the B-factor values of a novel protein.

Residue encoding scheme and normalization of B-factors

We encode the feature vector in our regression problem, X_i, representing a single residue in a protein, using sequence information in a window of width 15 centered on the residue[25]. We use two different encodings for protein residues: single-sequence and multiple-sequence. Each of these encodings represents a residue by a vector of length 21. The feature vector for a residue in a protein is composed by concatenating the fifteen vectors representing the residues in the window centered on it. In both encodings, the feature vector encoding a residue in a protein is a (21*15=) 315-dimensional vector.

In the single-sequence encoding, a residue is encoded as a vector of length 21. One vector element, corresponding to the residue type, is set to one, and the other twenty elements are set to zero. The first twenty elements in the vector each represent one of the 20 standard amino acids. The twenty-first element represents gaps, off-the-end of the sequence and non-standard amino acids.

The multiple-sequence encoding incorporates evolutionary information by encoding residues via their columns in a PSI-BLAST position-specific scoring matrix.[26] The matrices were obtained by querying the input protein using PSI-BLAST against the non-redundant protein sequence database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) with three rounds, masking coil-coiled and low-complexity regions[27]. The elements in PSI-BLAST position-specific scoring matrices are the log-odds values, which represent the log-likelihood of that particular residue substitution at that position in the template. To normalize the log-odds values, we divided them by 10 so that nearly all the values between –1.0 and 1.0. For masked residues, we encode the residue in the single-sequence format. For non-masked residues, the first 20 elements in the vector are the normalized log-odds values and the last one is assigned as zero.

The predicted variable in our regression problem, y_i, is the normalized B-factor of the C_α atom of the corresponding residue. We extract raw B-factor values from the protein's PDB entry[28] and normalize them using the method of Smith et al.[17] First, outliers are detected by a median-based method and removed from the input. Then, the sample mean ($\bar{y}$) and sample standard deviation (S) of data points are calculated. Finally, the raw B-factors, y_i', are normalized according to the following equation:

$$y_i = \frac{y_i' - \bar{y}}{S}, \; i = 1, \ldots, M. \tag{11}$$

Dataset preparation and accuracy measurement

We prepared 766 protein chains using PDB-REPRDB[29] and selected only the structures solved by X-ray crystallography with resolution ≤ 2.0 Å^2 and R-factor ≤ 0.2. We excluded protein chains shorter than 60 amino acids. No two chains were

included that have pair-wise identity more than 25%. The protein names can be found in Table 2 (supplementary material).

To perform 5-fold cross-validation tests, we divided this dataset into five groups, with groups having roughly equal numbers of protein sequences. One group in turn was chosen as the testing set, while the proteins in other groups were merged to form a training set.

To measure the performance of SVR methods, we calculated the CC between predicted and observed B-factors as given by

$$\rho = \frac{\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{M}(x_i - \bar{x})^2][\sum_{i=1}^{M}(y_i - \bar{y})^2]}} \tag{12}$$

where x_i and y_i are the experimental and predicted values of the B-factor of the ith C_α atom, and $\bar{x}$ and $\bar{y}$ are their corresponding sample means.

To further examine the performance, we chose a series of cutoff thresholds (to classify B-factor values. Residues are classified as flexible (rigid) if their B-factors are greater than (less than or equal to) the threshold. The overall accuracy is defined as the number of correctly predicted residues over the total number of residues. The sensitivity is defined as the ratio between the number of correctly predicted flexible residues and the number of flexible residues. Specificity is the ratio between the number of correctly predicted rigid residues and the number of rigid residues.

We also measure classification accuracy using Receiver Operating Characteristic (ROC) analysis.[30,31] ROC is a classic method in Signal Detection Theory[31] and was also used in medical diagnosis.[30] ROC plots, for every possible cutoff threshold, classification sensitivity as a function of one minus specificity,

should have higher sensitivity and specificity equates with better classification performance, so the highest and leftmost ROC curve represents the best-performing classification method.

Results

The distribution of normalized B-factors for the 191,474 C_α atoms in our dataset is similar that of datasets used by other authors in related work.[17,18] As shown in Fig 1, the distribution is skewed with mean and median values 0.03 and –0.23, respectively.

We tried four different combinations of input encoding, kernel function and regularization. These four "models" are described in detail in Table 1. Model 1 uses the single-sequence input encoding and a fourth-order polynomial kernel function. Models 2 and 3 use the single-sequence input encoding and radial basis function kernels. The RBF kernel performs better than the polynomial kernel, which was observed by our previous studies in the prediction of solvent accessible surface area.[24] Based on RBF kernel, the multiple-sequence input encoding is used to make model 4. We set the value of ε (Eq. 2) to 0.01 when training each model using SVR.

Among the four models, Model 4 gives the most accurate predictions of B-factors. For Model 4, the cross-validated correlation coefficient is 0.53. The equivalent results for models 1, 2 and 3 are 0.40, 0.43 and 0.44, respectively. As expected, the evolutionary information contained in the multiple-sequence input encoding significantly improves the prediction of B-factors.

The distribution of the correlation coefficients for individual protein chains (computed using model 4) is shown in Fig 2. More than half of the proteins have correlation coefficients greater than 0.56. The correlations for individual proteins ranged from –0.20 to 0.89 and their values are given in Table 2 (supplementary material). Among them, only four protein chains are predicted with negative correlation coefficients.

Each of the four models can also be used to classify residues into two classes (rigid and flexible). We selected the thresholds from –2.0 to 2.0 with a step size of

0.1 and calculated the corresponding accuracies. Fig. 3 shows that model 4 achieves equal or better classification accuracy than the other models regardless of the B-factor threshold used to define the two classes. For all choices of threshold, the *relative* classification accuracies of the models are always the same (from best to worst): model 4, model 3, model 2 and model 1. It is also apparent from Fig. 3 that the B-factor threshold used to define the two classes strongly affects the *absolute* classification accuracy. If the threshold is set at the mean value (0.03) or the median value (-0.23), the corresponding model 4 classification accuracies are 71.3% and 69.5%, respectively.

As can be seen from the ROC plots in Fig. 4, when we set the B-factor threshold to its mean value (0.03), model 4 dominates the other models. That is to say, with this definition of flexible/rigid residues, model 4 has better sensitivity for any choice of specificity when compared to the other models. This same result holds for all choices of threshold we tried (data not shown).

Fig. 5 illustrates the meanings of different values of the correlation coefficient. It shows the predicted and experimental flexible profiles for two particular proteins: cytokine 1KXG (chain A) and binding protein 1PUC. The correlation coefficients for 1KXG and 1PUC are 0.54 and 0.78, respectively.

The largest (absolute) errors in the representative flexible profiles shown in Fig. 5 occur for residues with the largest (absolute value) experimental B-factors. This is a general characteristic of our method, as can be seen in Fig. 6. In that figure we plot the absolute value of the difference between predicted and experimental B-factors for each residue in our test sets versus the experimental B-factor of the residue.

A probable explanation for the lower accuracy of the method on residues with very large and very small B-factors is the fact that these types of residues are much

less common in the training sets. It is a well-known property of machine learning algorithms (such as SVR) that predictive accuracy is often strongly correlated with the amount of training data available.

Discussion

The accuracy of the prediction of B-factors based on protein sequences or structures has an upper limit due to the noise in B-factor experimental data. Measured B-factors not only reflect the authentic fluctuation and static, dynamic and lattice disorders, but they also depend on refinement methods, stages and temperatures.[32-34] The presence of ligands may also impact the B-factor values of active sites. Some non-biological contributions such as crystal contacts are also the sources of noise. However, a careful examination by Radivojac et al.[12] showed that the B-factor distributions of homologous protein sequences are highly correlated (average correlation coefficient 0.80). This coefficient may be regarded as a prediction limit.

Several sequence-based B-factor prediction methods were compared by Radivojac et al.[12] using a smaller training dataset (290 protein chains). The method of Vihinen el al.[3] achieved a correlation coefficient of 0.32. The best method of based on neural networks and using multiple-sequence encoding as input, achieved a correlation coefficient of 0.43.[12] Based on our large dataset of high-resolution proteins, our best method has a correlation coefficient of 0.53, and more than half of the proteins are predicted with correlation coefficients greater than 0.56. This improvement may be partially due to our larger dataset (766 protein chains). Radivojac et al.[12] also report classification accuracies, but these are not directly comparable with our results because they use a different B-factor normalization method.

Different models have been used to predict B-factor distribution based on protein atomic coordinates. The translation liberation screw model[35-37] simplified the protein as a rigid body with movement along translation, liberation and screw axes. Considering only the liberation movement and defining B-factors as the squared distances of C_α atoms from the protein mass centroid, Kundu et al.[14] achieved an average correlation coefficient of 0.52 on a dataset of 113 high-resolution proteins. This result is comparable with our result (0.53) that is obtained from protein sequences.

More complicated structure-based analytical methods may predict B-factor distribution more accurately. The Gaussian network model (GNM)[38-41] regards a protein as an elastic network of C_α atoms that fluctuate about their mean positions. The fluctuations are assumed to be isotropic (without directional preferences) and also to obey Gaussian distributions. The mean-squared displacements of C_α atoms are used to define their B-factor values. Environmental parameters such as ligands and crystal contacts may be fed into the model and yield different B-factors values. Kundu et al.[14] found that GNM using C_α atom coordinates in an isolated molecule could achieve a correlation coefficient of 0.59 with the experimental B-factors. This correlation coefficient was improved to 0.66 if the atoms involved in crystal contacts had been excluded from calculation. In addition, a correlation coefficient of 0.66 could also be achieved if all neighbouring molecules that may make crystal contacts were considered in the model.

Our SVR approach is trained and tested on primary protein chains only. Therefore, some portions of B-factors, which are impacted by protein chain-chain contacts, protein-ligand contacts and molecule crystal contacts (non-biological contacts), cannot be well predicted by our method. The purpose of this study is to find

a mapping function between protein sequences and their B-factor distributions using a machine learning approach. In this work, our method is trained using B-factor data of molecules in crystals, which may not reflect the real situation of molecules in solution. Were adequate training data available, we could train our method using biological B-factor data. This is a good topic for future work.

The advantage of our method is that it does not require any structural information about the input protein. This is very useful when one wants to obtain some information about protein flexibility when only its sequence is known. This information is useful for investigating protein function, since active sites are often located in flexible loop regions. Furthermore, prediction of protein flexibility profiles complements the prediction of other protein properties such as disordered regions[12,42,43] and surface residues.[25,44,45]

Conclusion

We provide an SVR approach for prediction of protein B-factor distributions (flexibility profiles) from sequence data alone. Because our method requires only the sequence of the protein as input, it is applicable to the large numbers of proteins whose structures are not yet known. Our best method, using multiple sequence alignment input encoding and a radial basis function kernel, greatly improves on the prediction accuracy of previous methods (CC of 0.53 compared with 0.43). In addition, our method achieves a correlation coefficient of 0.56 or better more than half of the time.

Acknowledgment

The work was supported by funds from the Australian Research Council (ARC). The computer simulations were performed at the High Performance Computing Facility at the University of Queensland. T.L.B. is supported by a DETYA grant. R.D.T. is supported by a National Health and Medical Research Council of Australia R. Douglas Wright Career Development Award.

References

1. Daniel RM, Dunn RV, Finney JL, Smith JC. The role of dynamics in enzyme activity. Bioinformatics 2003;32:69-92.
2. Karplus PA, Schulz GE. Prediction of Chain Flexibility in Proteins - a Tool for the Selection of Peptide Antigens. Naturwissenschaften 1985;72:212-213.
3. Vihinen M, Torkkila E, Riikonen P. Accuracy of Protein Flexibility Predictions. Proteins 1994;19:141-149.
4. Vihinen M. Relationship of Protein Flexibility to Thermostability. Protein Eng 1987;1:477-480.
5. Parthasarathy S, Murthy MRN. Protein thermal stability: insights from atomic displacement parameters (B values). Protein Eng 2000;13:9-13.
6. Carugo O, Argos P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. Proteins 1998;31:201-213.
7. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109-114.
8. Mohan S, Sinha N, Smith-Gill J. Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: An insight into the molecular basis of antibody cross-reactivity and specificity. Biophys J 2003;85:3221-3236.
9. Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Eng 1997;10:777-787.

10. Eyal E, Najmanovich R, Edelman M, Sobolev V. Protein side-chain rearrangement in regions of point mutations. Proteins-Structure Function and Genetics 2003;50:272-282.

11. Altman R, Hughes C, Zhao D, Jardetsky O. Compositional characteristics of relatively disordered regions in proteins. Prot Pept Lett 1994;1:120-127.

12. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. Protein Sci 2004;13:71-80.

13. Navizet I, Lavery R, Jernigan RL. Myosin flexibility: Structure domains and collective vibrations. Proteins 2004;54:384-393.

14. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of proteins in crystals: Comparison of experiment with simple models. Biophys J 2002;83:723-732.

15. Ming D, Kong YF, Lambert MA, Huang Z, Ma JP. How to describe protein motion without amino acid sequence and atomic coordinates. Proc Natl Acad Sci U S A 2002;99:8620-8625.

16. Tronrud DE. Knowledge-based B-factor restraints for the refinement of proteins. J Appl Crystallogr 1996;29:100-104.

17. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12:1060-1072.

18. Parthasarathy S, Murthy MRN. Analysis of temperature factor distribution in high-resolution protein structures. Protein Sci 1997;6:2561-2567.

19. Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, New York; 2000.

20. Smola A, Scholkopf B. A tutorial on support vector regression. NeuroCOLT Technical Report Serie, NC-TR-1998-030, http://wwwneurocoltcom; 1998.

21. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Mozer MC, Jordan MI, Petsche T, editors: MIT Press,Cambridge, MA.; 1997. 155-161 p.

22. Myasnikova E, Samsonova A, Samsonova M, Reinitz J. Support vector regression applied to the determination of the developmental age of a Drosophila embryo from its segmentation gene expression patterns. Bioinformatics 2002;18:S87-S95.

23. Segal MR, Dahlquist KD, Cionklin BR. Regression approaches for Microarray data analysis. J Comput Biol 2003;10:961-980.

24. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;In press.

25. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216-226.

26. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-3402.

27. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.

28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.

29. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 2003;31:492-493.

30. Centor RM. Signal detectability: The use of roc curves and their analyses. Med Decis Making 1991;11:102-106.

31. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press; 1982.

32. Frauenfelder H, Petsko GA, Tsernoglou D. Temperaure-dependent X-ray diffraction as a probe of protein structural dynamics. Nature 1979;280:558-563.

33. Ringe D, Petsko GA. Study of protein dynamics by X-ray diffraction. Methods Enzymol 1986;131:389-433.

34. Carugo O, Argos P. Reliability of atomic displacement parameters in protein crystal structures. Acta Crystallogr D 1999;55:473-478.

35. Schomaker V, Trueblood KN. On the rigid-body motions of molecules in crystals. Acta Crystallogr B 1968;24:63-76.

36. Sternberg MJE, Grace DEP, Phillips DC. Dynamic information from protein crystallography: an analysis of temperature factors from refinement of the hen egg white lysozyme structure. J Mol Biol 1979;130:231-253.

37. Kuriyan J, Weis WI. Rigid protein motion as a model for crystallographic temperature factors. Proc Nat Acad SCI USA 1991;88:2273-2277.

38. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173-181.

39. Haliloglu T, Bahar I. Gaussian dynamics of folded proteins. Phys Rev Lett 1997;79:3090-3093.

40. Bahar I, Atilgan AR, Demirel MC, Erman B. Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. Phys Rev Lett 1998;80:2733-2736.

41. Haliloglu T, Bahar I. Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with x-ray diffraction and NMR relaxation data. Proteins 1999;37:654-667.

42. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53:566-572.

43. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53:573-578.

44. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566-570.

45. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629-635.

Table 1. Models: sequence coding schemes, kernel functions and control parameters.

Model	Encoding scheme	Kernel function	Parameters
1	Single-sequence	$K(X_i, X_j) = (X_i \cdot X_j + 1)^n$	n=4; C=0.001
2	Single-sequence	$K(X_i, X_j) = \exp(-\gamma \left\| X_i - X_j \right\|^2)$	γ =0.1; C=0.7
3	Single-sequence	$K(X_i, X_j) = \exp(-\gamma \left\| X_i - X_j \right\|^2)$	γ =0.01; C=2.0
4	Multiple-sequence	$K(X_i, X_j) = \exp(-\gamma \left\| X_i - X_j \right\|^2)$	γ =0.01; C=2.0

Figure legends:

Fig. 1. Distribution of normalized B-factors. Each point shows the percentage of residues with B-factors in a bin of size of 0.2 among the total residues in the 766 protein chains in the dataset.

Fig. 2. The distribution of correlation coefficients between predicted and experimental B-factors. Each point shows the percentage of proteins in the dataset with correlation coefficients in a bin of size of 0.1.

Fig. 3. Prediction accuracies versus cutoff thresholds for different models.

Fig. 4. ROC analysis of different models when the B-factor classification threshold is set as 0.03.

Fig. 5. Prediction results for protein 1KXG (chain A) and 1PUC. The dashed lines represent B-factors predicted by SVR and the solid lines represent the measured B-factors. The correlation coefficients for 1KXG and 1PUC are 0.54 and 0.78, respectively.

Fig. 6. Absolute prediction error versus experimental B-factors. Lower and higher B-factors are predicted with greater (absolute) errors.

Fig. 1

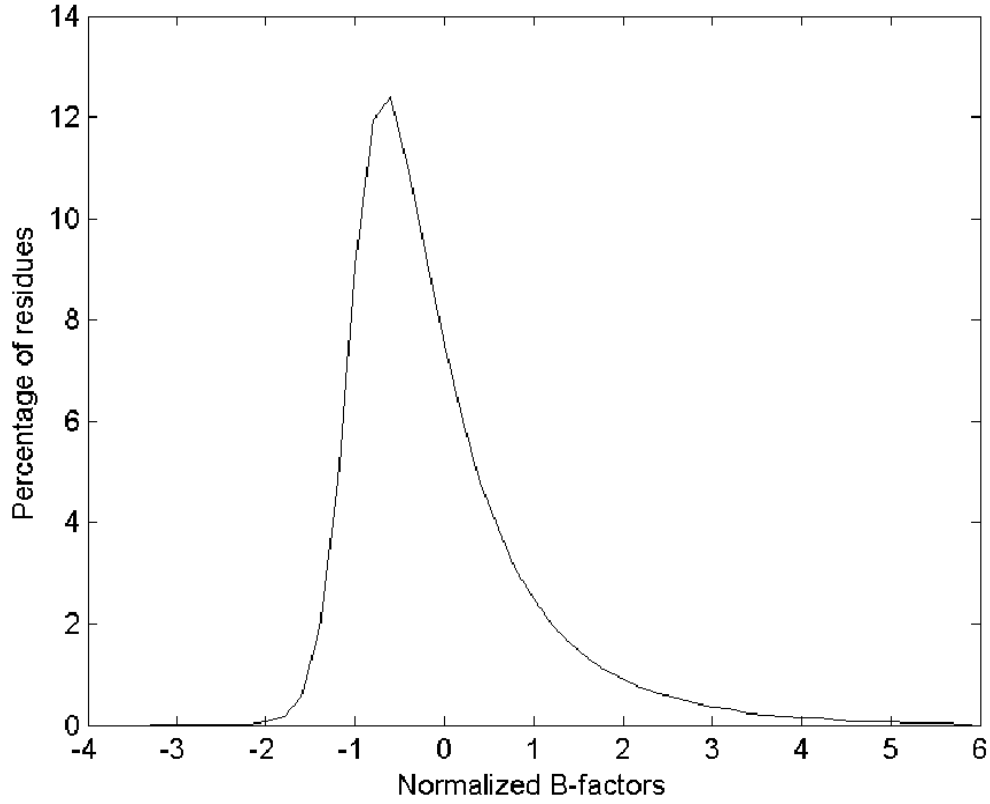

Percentage of residues
Normalized B-factors
0
2
4
6
8
10
12
14
-4
-3
-2
-1
0
1
2
3
4
5
6

Fig. 2

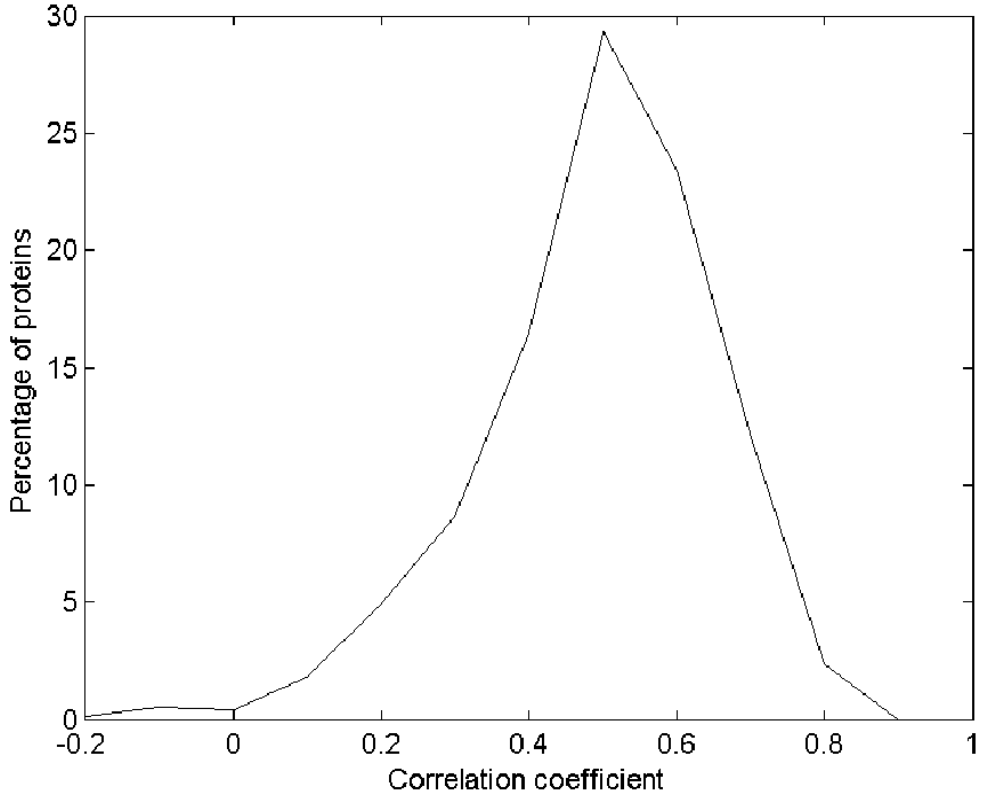

30
25
20
15
10
5
0
Percentage of proteins
-0.2
0
0.2
0.4
0.6
0.8
1
Correlation coefficient

Fig. 3

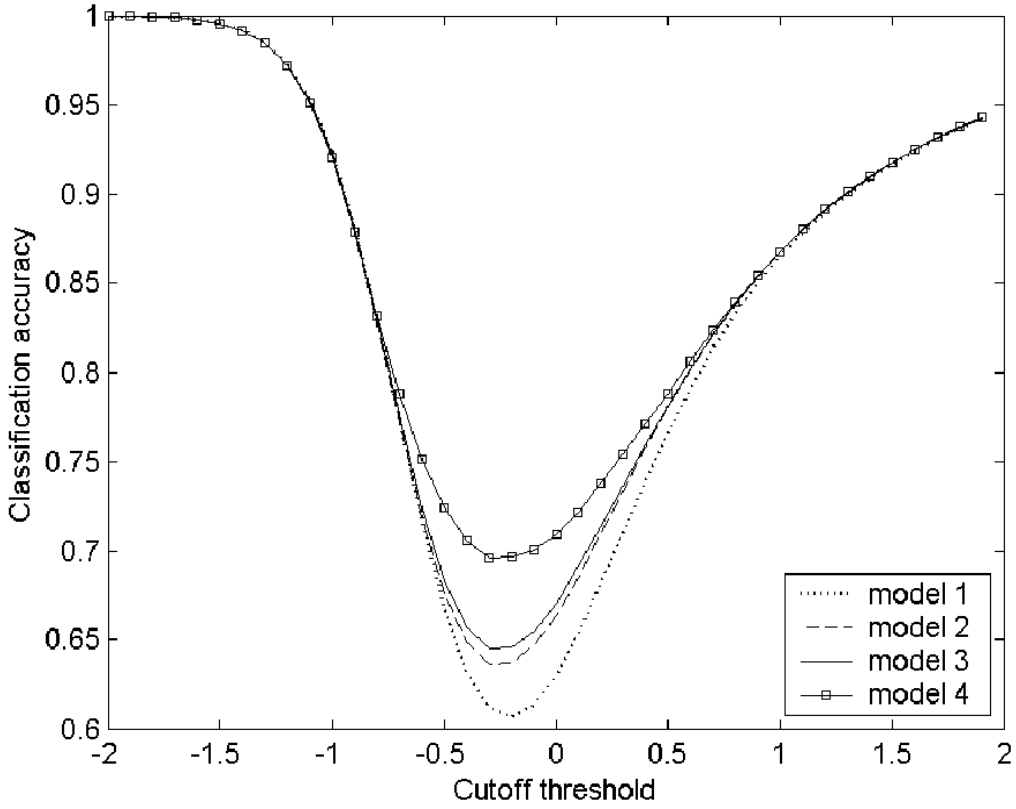

Classification accuracy
Cutoff threshold
model 1
model 2
model 3
model 4
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Fig. 4

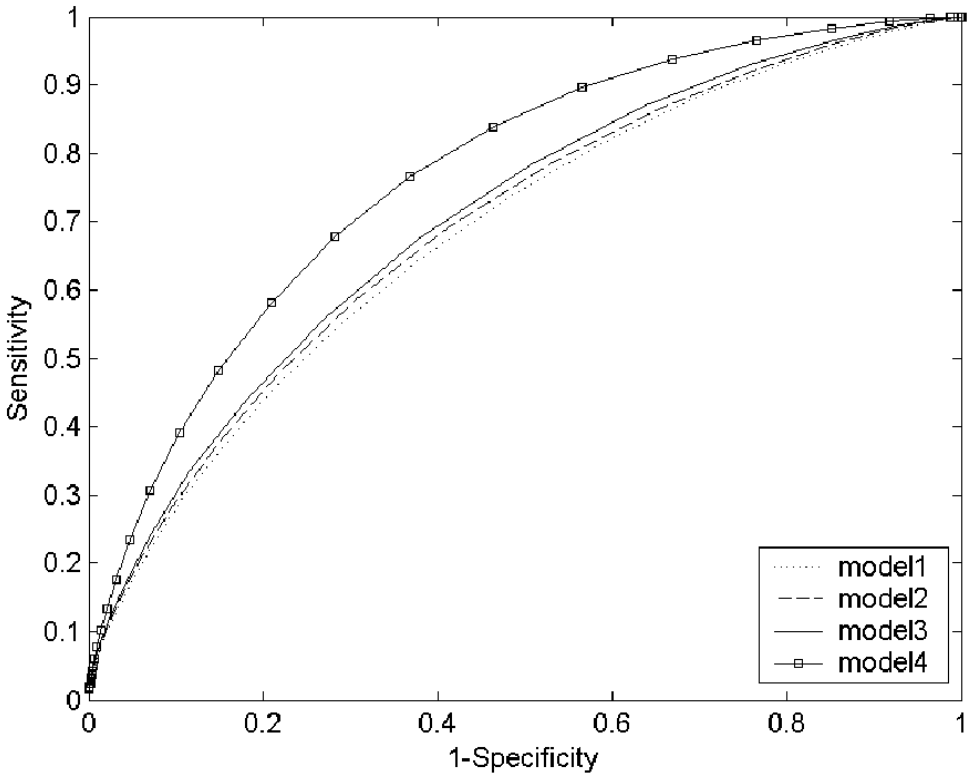
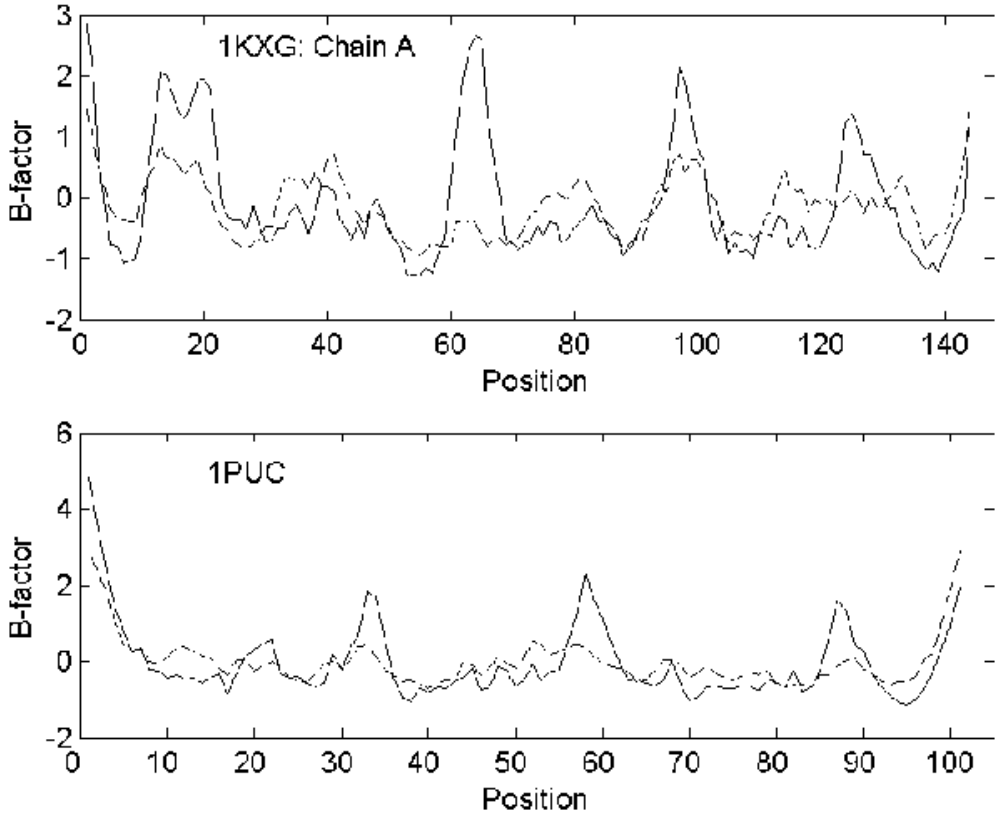

Sensitivity
1-Specificity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
model1
model2
model3
model4

Fig. 5



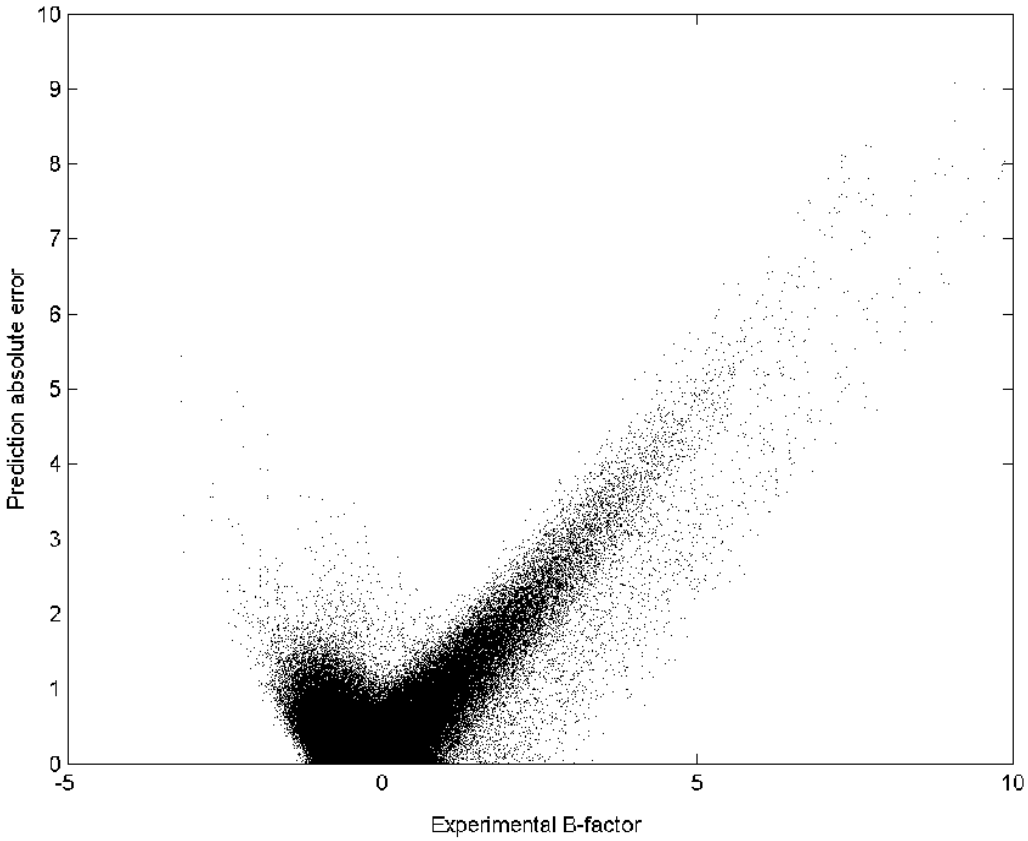


Fig. 6