

Prediction of temperature factors from protein sequence

Shrihari Sonavane^{1*}, Ashok A Jaybhaye² & Ajaykumar G Jadhav¹

¹Department of Microbiology, Institute of Science, Caves Roed, Aurangabad, India-431004; ²Maharashtra Hybrid Seeds Company Limited Jalna-Aurangabad Rd, Jalna, India-431004; Shrihari Sonavane - Email: shriharry@gmail.com; Ph: +91-8275267549;

*Corresponding author

Received January 06, 2013; Accepted January 10, 2013; Published February 06, 2013

Abstract:

Protein flexibility is useful in structural and functional aspect of proteins. We have analyzed the local primary protein sequence features that in combination can predict the B-value of amino acid residues directly from the protein sequence. We have also analyzed the distribution of B-value in different regions of protein three dimensional structures. On an average, the normalized B-value decreases by 0.1055 with every 0.5 Å increase in the distance of the residue from protein surface. The residues in the loop regions have higher B-values as compared to the residues present in other regular secondary structural elements. Buried residues which are present in the protein core are more rigid (lower B-values) than the residues present on the protein surface. Similarly, the hydrophobic residues which tend to be present in the protein core have lower average B-value than the polar residues. Finally, we have proposed the method based on Support Vector Regression (SVR) to predict the B-value from protein primary sequence. Our result shows that, the SVR model achieved the correlation coefficient of 0.47 which is comparable to existing methods.

Keywords: B-value, Protein flexibility, Support Vector Regression, Sliding window approach, Protein dynamics.

Background:

Protein structures are dynamic molecules which are in constant motion. The protein motion or flexibility is highly correlated with various biological processes such as molecular recognition and catalytic activity. In theory the protein flexibility is studied by computational models of structure dynamics, atomic normal mode analysis (NMA) and by simulations of molecular motion, while experimentally it is probed by techniques such as nuclear magnetic resonance relaxation times, incoherent neutron scattering and X-ray structure B-values (Temperature factor or B- values) [1-5]. B-value reported in experimental atomic-resolution structures represents the decrease of intensity in diffraction due to the dynamic disorder caused by the temperature-dependent vibration of the atoms and the static disorder, which is related to the orientation of the molecule [6]. Generally the B-values of C-alpha atoms are used to study the motion of the backbone. It is clear that B-value data of crystal structures are influenced by a number of disparate variables and importantly on the particular refinement procedures [6]. As

a result, measured B-values in different known structures may be on different scales; normalized B-values are used to compare the B-values of different protein structures [7-10]. Earlier works aimed at analysis of B-values obtained from protein X-ray crystal structure have shown that the prediction of B-values can help to unravel protein function along with protein thermal stability, [11, 12] side-chain mobility with conformation, [13, 14] protein disordered regions [15, 16] and investigating protein dynamics [17].

Statistical analysis has formed a good component of the research efforts aimed at understanding protein flexibility, owing to the enormous diversity and complexity. A variety of approaches have been proposed by different groups to predict the B-values, mainly dealing with the so called "classification problem". In classification problem the amino acid residues are classified into two states as rigid or flexible on the basis of an arbitrary B-value cutoff threshold. [7, 8, 16]. In contrast, method proposed by Yuan *et al.* [18] predicts a series of real values

representing a protein sequence (also regarded as the B-value profile) only using the single-sequence and multiple-sequence (multiple sequence alignment) features. This approach achieves the correlation coefficient of 0.53 which is better than the previous approaches.

In this study, we have analyzed the B-values from a non-redundant data set of high-resolution structures to look into the variations in B-value of residues due to their biochemical properties as well as the environment around them in protein three dimensional structures. We have also proposed the method using Support Vector Regression (SVR) to predict B-value from protein sequence.

Methodology:

Dataset

A non-redundant set of protein chains was selected from the Protein Data Bank [19] (PDB) located at Research Collaboratory for Structural Bioinformatics (RCSB) having resolution better than 2.0 Å. These entries were submitted to the PISCES server [20] to identify proteins with low sequence similarity. Following are the input parameters used for culling: sequence percentage identity 25%, resolution 2.0 Å, R-factor 0.2, sequence length 60-10000 amino acids. Final dataset consist only monomeric protein structures. The multimeric proteins are removed for the sake of simplicity. Protein chains containing fragments and missing C-alpha atom coordinates are also removed. The resulting data set contained 240 protein chains, and the list is provided in Table S1 (see supplementary material). High resolution protein structures containing B-values data were taken from Protein Data Bank (PDB) [19]. Secondary structure assignments were made using the DSSP program. [21]. Amino acid residues are grouped into one of the three secondary structural states as helix, stand and the rest, termed 'Others'.

Calculation of normalized B-value

As B-values from different structures are on different scales, the raw B-values from the Protein Data Bank (PDB) for each protein are normalized for reasonable comparison by the following equation, (see supplementary material for Formula 1).

Identification of clefts and cavities

The CASTp (Computed Atlas of Surface Topography of proteins) server located at <http://sts.bioengr.uic.edu/castp/> was used to identify clefts and cavities for each protein structure [22]. CASTp provides a full description of protein pockets and cavities including volume, surface area, protein atoms that line the concavity and features of pocket mouth(s) including identification of mouth atoms as well as measurement of mouth area and circumference. The default probe radius of 1.4 Å has been used for our calculations.

Calculation of protein accessible surface area and residue distance from the protein surface

The calculation of protein accessible surface area (ASA) was done by the program naccess, [23] which implements the algorithm of Lee and Richards [24]. All the atoms having accessible surface area > 0 Å² are considered as the surface atoms. The Euclidian distance of the C-alpha atom of the residue from the nearest surface atom is considered as the effective distance of the residue from the protein surface.

Support Vector Machines implementation

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 9(3):134 -140 (2013)

Support Vector Machines (SVMs) are a class of machine learning algorithms which can perform pattern recognition and regression [25, 26]. SVMs can handle noise and large datasets very effectively. SVM non-linearly transforms the original input space into a higher dimensional feature space by means of kernel functions [27, 28]. The freely downloadable LIBSVM package was used for the implementation of SVM with the nu-SVR SVM type (SVM type for regression) and the widely used Radial Basis Function (RBF) kernel [29]. Out of 240 protein structures considered in this study 70% were randomly picked for creating the training dataset consisting of 168 cases. The remaining 30% were used as test set (72 cases). The SVR model was also tested on 766 cases from previous studies. All the attributes in training and test datasets were scaled in the range of -1 to 1.

Parameter Selection and encoding

Altogether thirty three sequence features were used as attributes for implementing the SVR model. The attributes were amino acid type (20 amino acids), amino acid class based on biochemical features (4 classes) and the amino acid class as classified by Saha et. al. [30] based on the similarity of the environment of each amino acid residue in protein structures (9 classes). The input feature vectors of SVR are encoded for each residue in the primary protein sequence based on the sliding window approach centered at middle residue except near the N- and C-termini. Different window sizes (from 3 to 21 residues) are tested and the optimized window size is set to 15 residues, which gives the better results. Each residue is represented by a vector of length 36; 21 for amino acid type (The first 20 elements in the vector each represent one of the 20 standard amino acids and the twenty first element represents non-standard amino acids.), 5 for amino acid class based on biochemical features and 10 for amino acid class based on the similarity of the environment of each amino acid residue in protein structures [30]. Therefore, each residue site in the protein sequence is encoded by 36*15=540 dimensional vector.

SVM optimization

The nu-SVR parameter C and the kernel parameters γ had been optimized using repeated grid search and 5-fold cross validation. The e-epsilon parameter is set to 0.01 as it gives the best results. In 5-fold cross-validation, the training dataset was split into 5 subsets, where one of the subsets was used as the test set while the other 4 subsets were used for training the SVR model. The trained model was tested using the test set. The process was repeated 5 times using a different subset for testing, thereby ensuring that all subsets were used for both training and testing.

Performance measure

We have calculated the Pearson correlation coefficient (CC) between the theoretical and experimental B-values to verify the performance of SVR method, as given by (see supplementary material for Formula 2).

Discussion:

B-value profile for residues in different regions of protein tertiary structure

The B-value distributions of the residues vary depending on their positions in the protein tertiary structure. Dataset considered in this study consists of 63771 residues across 240

protein structures. The normalized B-value ranges from -2.29 to 11.96. **Figure 1** shows the frequency distribution of B-values for amino acid residues in overall dataset and present in different position in protein tertiary structure. The frequency distribution of normalized B-value in overall dataset (**Figure 1A**) shows the skewed distribution with the flat peak having two maximas at -0.5 and 1.

Exposed residues have higher B-value than the others which reside in protein core (**Figure 1B**). The respective average B-values for the buried, intermediate and exposed residues are -0.417, 0.210 and 1.092, indicating that the buried residues are on average, less flexible than the exposed residues. The overall shapes of the B-value distributions in cavities and pockets (**Figure 1C**) are similar, except the distribution of B-values for

residues lining the pockets is relatively flat. This may be due to the fact that cavity lining residues are not solvent accessible while the residues lining the pockets are solvent accessible.

It (**Figure 1D**) shows the frequency distribution of B-values for amino acid residues in α -helix, β -strand and others. The β -sheet and 'Others' distributions have maximum values at -0.05 and 0, respectively. Whereas, the B-value distribution of the residues in the helical states shows skewed distribution with the flat peak (maximas at -0.5 and 0). The average B-values for the helical, sheet and 'Other' residues are -0.11, -0.4 and 0.31, respectively, indicating that the coil residues, in general, have larger B-values and therefore are more flexible than helical and sheet residues.

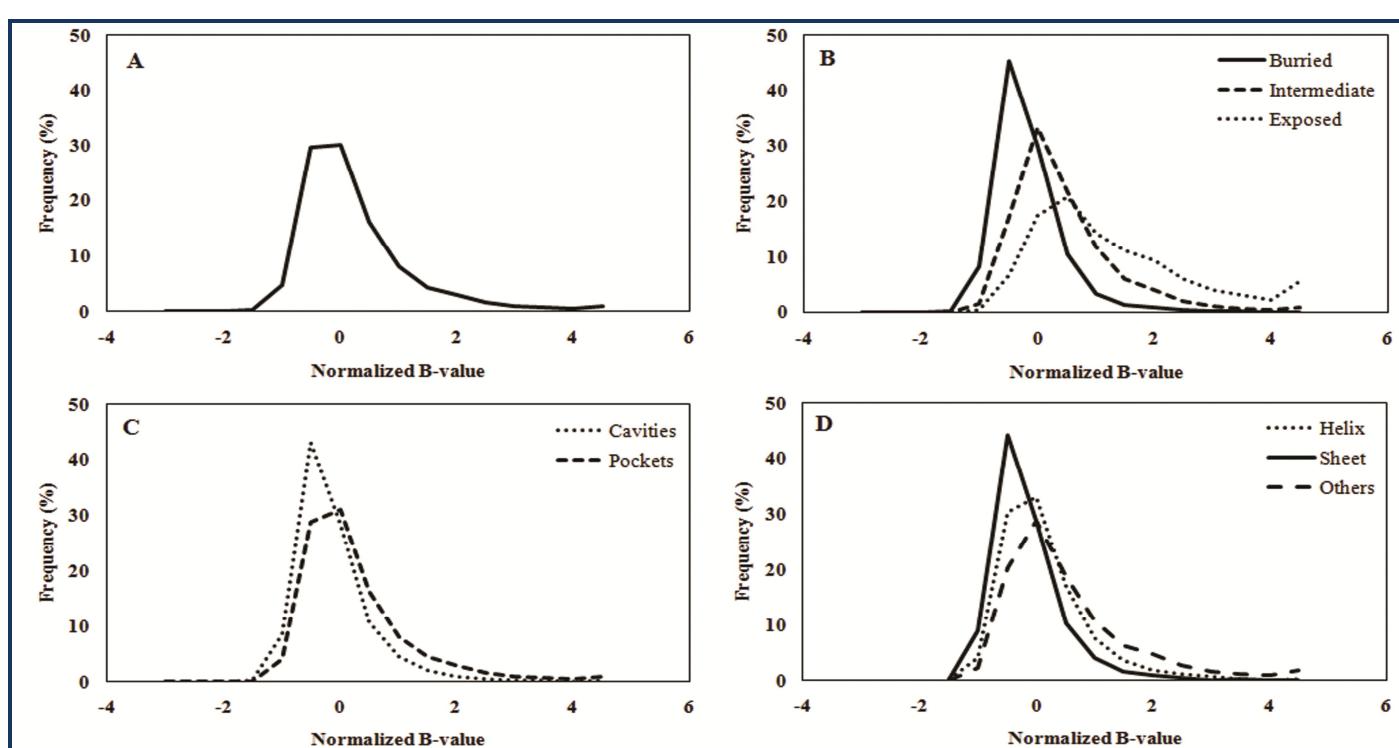


Figure 1: Frequency distribution of normalized B-value in bins of 0.5. The plot showing (**A**) overall distribution; (**B**) distribution in solvent accessible and buried residues; (**C**) residues lining the cavities and the pockets, and; (**D**) the residues associated with the different secondary structural elements in the whole dataset.

Residue distance from protein surface and B-value

The accessible surface area of protein and the distance of the residue from the protein surface are calculated as described in the "Methods and Materials" section. There is a considerable variation in the normalized B-value of the residue as a function of its distance from the protein surface (burial). To discern any underlying trend we averaged the normalized B-value in a particular distance bin (size 0.5 Å) (Figure 2) and based on the average numbers one can derive a linear relationship. The plot of the normalized B-value of the residue against its distance from the protein surface (**Figure 2**) shows that the B-value decreases as the residue distance from its surface increases. The R^2 value for the straight line which fits on this data is 0.73 (equation $y = -0.2111x + 0.3434$). The normalized B-value decreases by 0.1055 with every 0.5 Å increase in the distance of the residue from protein surface.

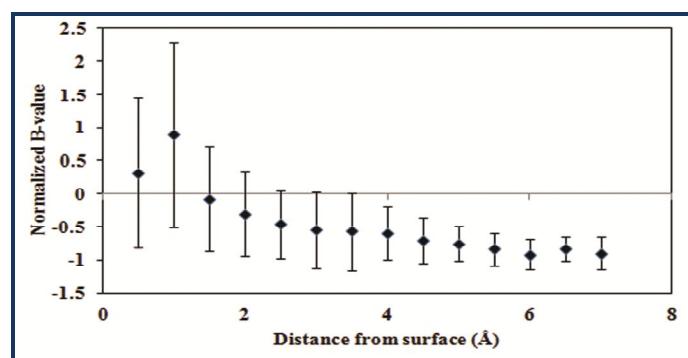


Figure 2: Plot of relationship between distance of C-alpha atoms from the protein surface and the normalized B-value. The average of all the values in a bin (size 0.5 Å) along the x axis is shown at the upper limit of each bin; while the y value

corresponds to the mean of their normalized B-values (the vertical bars represent the standard deviations)

Dependence of B-value on amino acids types

Amino acids on the basis of their biochemical characteristics are divided into 3 types as polar charged, polar uncharged and non polar hydrophobic amino acids. The frequency distribution of normalized B-values for each type is shown in (Figure 3). The distribution of normalized B-value in polar uncharged residues is similar to that of the overall distribution in the dataset, while the distribution of normalized B-value in non polar hydrophobic and polar charged is similar to the one in buried and intermediate surface residues, respectively.

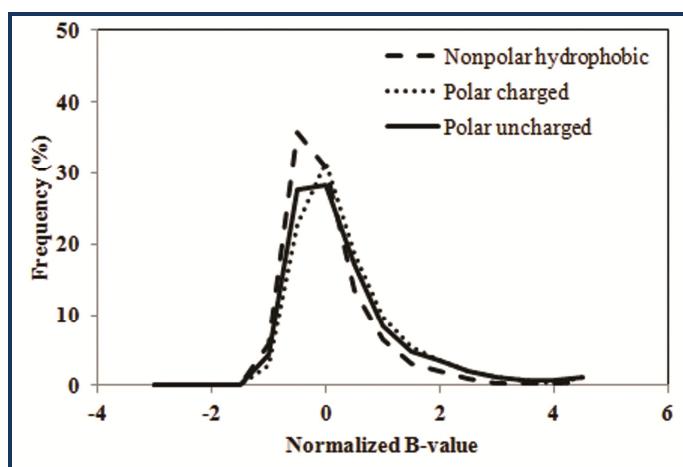


Figure 3: Frequency distribution of normalized B-value in polar charged, polar uncharged and non polar hydrophobic amino acids.

Dependence of residue B-value on protein primary sequence

The frequency distribution of B-value for each amino acid when the previous position (in protein primary sequence) is occupied by a particular amino acid (out of 20 amino acids) is shown in **Figure S1 (see supplementary material)**. The plot shows how particular amino acid residue affects the B-value of the next amino acid residue in the protein primary sequence, like if valine comes after cysteine in protein primary sequence then ~61% times its normalized B-value falls in bin -0.5 (i.e. in between -1.0 to -0.5). Similarly glutamine normalized B-value falls 60% times in between -0.5 and 0.

SVR model and prediction of B-value

The SVR model was designed using the RBF kernel and the 33 sequence feature input encoding as discussed in "Materials and Methods". We set the value of ϵ to 0.01 and p to 0.01 while training the model using SVR. The optimized values for the regularization parameter C and kernel parameter γ are listed in **Table 1 (see the supplementary material)**. We tried eight different window sizes to build the SVR model. Window size of 15 residues gives the best correlation coefficient (CC). Parthasarathy & Murthy [9] analyzed temperature factor distribution in high-resolution protein structures. The frequency distribution of normalized B-value in this study would match with Parthasarathy and Murthy's work. Our results show the skewed distribution with the flat peak. We have used the larger dataset created using more stringent criteria and it contains protein chains having length greater

than 60 amino acids. Nevertheless, Figure 1B indicates that the B-value distribution is flatter and more skewed towards the positive side if we go from buried to surface residues. It has been noted that long loops tend to be more flexible than regular secondary structures such as helices and strands. Moreover, hydrophobic residues, which are buried, tend to be more rigid whereas charged residues tend to be more flexible. These observations match well with the results obtained in this study (Figure 1, 3).

Proteins are not exactly spherical with smooth surface. Hence, a protein structural descriptor, the distance of an atom or residue from protein surface (or the depth) has been proposed to be the distance between itself (residue or atom) and the nearest water molecule or protein surface, to measure the extent to which it is buried [31, 32]. This parameter is useful in analyzing amide hydrogen/deuterium exchange rates in nuclear magnetic resonance (NMR) experiments [31]. Predicting protein active sites [33], and improving protein structural alignments [34]. Previous works have also shown that the average atomic fluctuation is linearly related to the square of the atomic distance from the center of mass of the protein which in turn is used to compute the B-value [35]. In contrast, if we look from the protein surface to core, the normalized B-value decreases by 0.1055 with every 0.5 Å increase in the distance of the residue from protein surface.

If structure-sequence relationships exist on a continuum, then one would expect to observe change in the B-value of amino acid based on the presence of particular amino acid at the nearby position in the protein primary sequence. The difference in the B-value distribution of amino acid based on the presence of particular amino acid at the presiding position in protein primary sequence as discussed in the results section – indicate that the protein primary sequence information can be useful in predicting the B-value. To investigate the amount of information required to optimally predict the correct B-value we have tried different window sizes of 7 to 21 residues in length. The window size of 15 residues has given the optimal results. Finally, we have built the SVR model using sliding window (window size of 15 residues) approach. The model achieved an average CC of 0.47, which is comparable with the other similar approaches. Different groups have used different datasets and methods to predict the B-values from protein sequence. As such, the direct comparison can't be an appropriate alternative in this case. Keeping this in mind our model has performed well on the test dataset as compared to other methods. Yuan et. al. [18] reported an average CC of 0.53 on the dataset of 786 pdb chains, but in their study they have not mentioned about the training and test dataset. Since the accuracy obtained on the whole dataset rather than dividing it in training and test (unseen) dataset is subject to be changed in case of future events. The method and the approach used in this study are more appropriate as the model has been tested on unseen test dataset.

Conclusion:

We have compiled a diverse data set using more stringent criteria and analyzed the usefulness of the local sequence features in identifying the residue B-value from protein primary sequence. A quantitative relationship has been derived for the

B-value of the residue and the residues depth from the protein surface. Finally, we have built the SVR model using the sliding window approach (window size of 15 residues) to predict the B-value of the amino acid residues from protein primary sequence and achieve results that are better or at least comparable to existing and reported methods.

References:

- [1] Williams RJ, *Eur J Biochem*. 1989 **183**: 479 [PMID: 2673776]
- [2] Sneddon SF & Brooks CL III, *University Press Oxford*. 1993 **115**
- [3] van Gunsteren WF et al. *Computer Phys Comm*. 1995 **91**: 305
- [4] LeMaster DM & Kushlan DM, *J Am Chem Soc*. 1996 **118**: 9255
- [5] Smith JC & Kneller GR, *Molecular Simulation*. 1993 **10**: 363
- [6] Creighton T, *W H Freeman New York*. 1993
- [7] Karplus PA & Schulz GE, *Naturwissenschaften*. 1985 **72**: 212
- [8] Vihinen M et al. *Proteins*. 1994 **19**: 141 [PMID: 8090708]
- [9] Parthasarathy S & Murthy MR, *Protein Sci*. 1997 **6**: 2561 [PMID: 9416605]
- [10] Smith DK et al. *Protein Sci*. 2003 **12**: 1060 [PMID: 12717028]
- [11] Vihinen M, *Protein Eng*. 1987 **1**: 477 [PMID: 3508295]
- [12] Parthasarathy S & Murthy MR, *Protein Eng*. 2000 **13**: 9 [PMID: 10679524]
- [13] Carugo O & Argos P, *Protein Eng*. 1997 **10**: 777 [PMID: 9342144]
- [14] Eyal E et al. *Proteins*. 2003 **50**: 272 [PMID: 12486721]
- [15] Altman R et al. *Prot Pept Lett*. 1994 **1**: 120
- [16] Radivojac P et al. *Protein Sci*. 2004 **13**: 71 [PMID: 14691223]
- [17] Navizet I et al. *Proteins*. 2004 **54**: 384 [PMID: 14747987]
- [18] Yuan Z et al. *Proteins*. 2005 **58**: 905 [PMID: 15645415]
- [19] Berman HM et al. *Nucleic Acids Res*. 2000 **28**: 235 [PMID: 10592235]
- [20] Wang G & Dunbrack RL Jr, *Nucleic Acids Res*. 2005 **33**: W94 [PMID: 15980589]
- [21] Kabsch W & Sander C, *Biopolymers*. 1983 **22**: 2577 [PMID: 6667333]
- [22] Binkowski TA et al. *Nucleic Acids Res*. 2003 **31**: 3352 [PMID: 12824325]
- [23] Hubbard SJ, A Program for Calculating Accessibilities. (Dept. of Biochemistry and Molecular Biology, Univ. College of London, London) 1992.
- [24] Lee B & Richards FM, *J Mol Biol*. 1971 **55**: 379 [PMID: 5551392]
- [25] Cortes C & Vapnik V, *Machine Learning*. 1995 **20**: 273
- [26] Muller KR et al. *IEEE Trans Neural Netw*. 2001 **12**: 181 [PMID: 18244377]
- [27] Gunn S, ISIS Technical Report ISIS-1-98. University of Southampton: Image Speech & Intelligent Systems Research Group; 1998
- [28] Kulkarni A et al. *Comput Chem Eng*. 2004 **28**: 311
- [29] <http://www.csie.ntu.edu.tw/~cjlin/libsvm> 2001
- [30] Saha RP et al. *J Proteome Res*. 2005 **4**: 1600 [PMID: 16212412]
- [31] Pedersen TG et al. *J Mol Biol*. 1991 **218**: 413 [PMID: 2010918]
- [32] Chakravarty S & Varadarajan R, *Structure*. 1999 **7**: 723 [PMID: 10425675]
- [33] Gutteridge A et al. *J Mol Biol*. 2003 **330**: 719 [PMID: 12850142]
- [34] Zhou H & Zhou Y, *Proteins*. 2004 **55**: 1005 [PMID: 15146497]
- [35] Shih CH et al. *Proteins*. 2007 **68**: 34 [PMID: 17436324]

Edited by P Kangueane

Citation: Sonavane et al. Bioinformation 9(3): 134-140 (2013)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

Supplementary material:

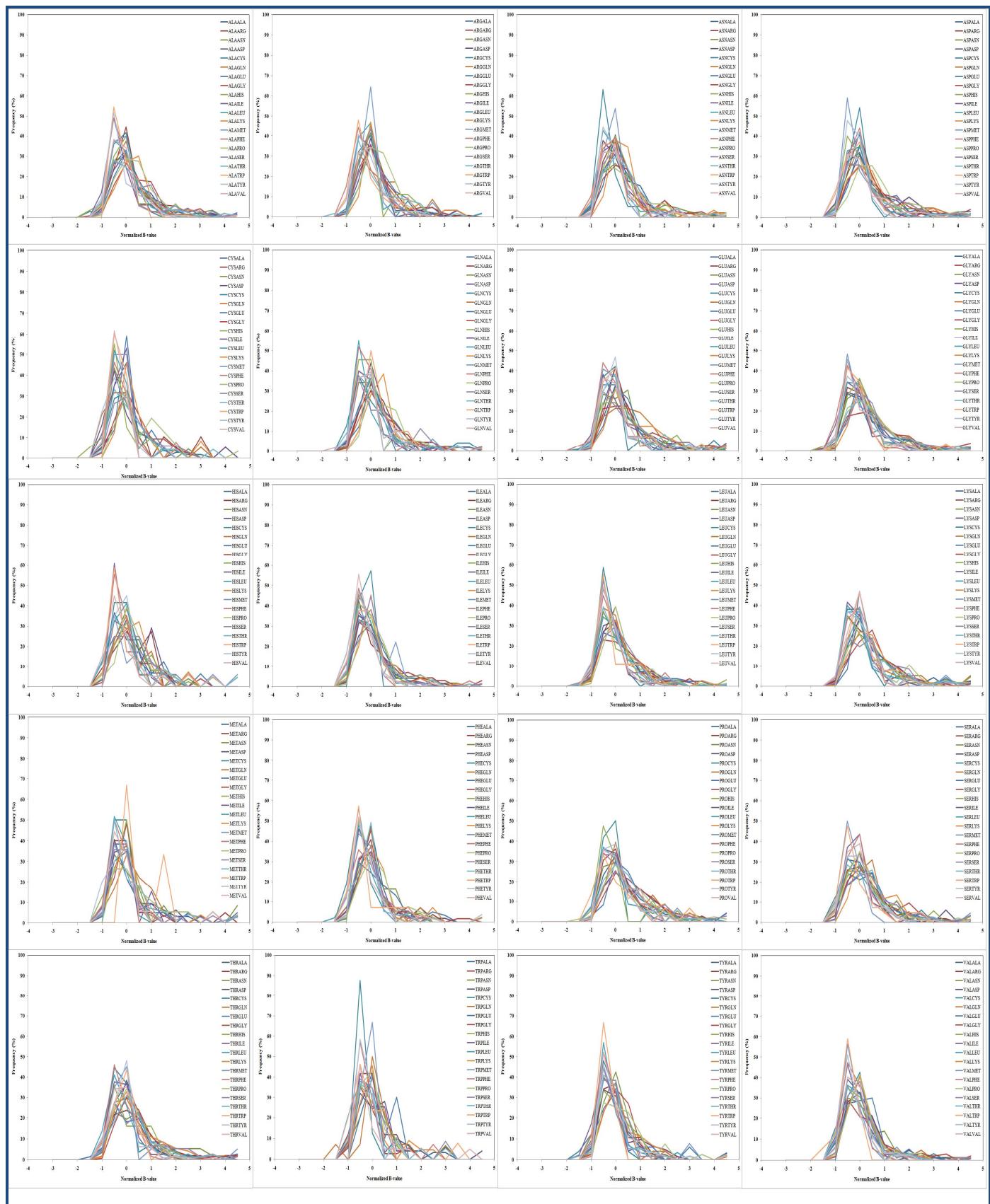


Figure S1: Normalized B-value frequency distribution for all 20 amino acids when the previous position in the primary protein sequence is occupied by particular amino acid

Methdology:

Calculation of normalized B-value

Formula 1:

As B-values from different structures are on different scales, the raw B-values from the Protein Data Bank (PDB) for each protein are normalized for reasonable comparison by the following equation,

$$\text{Normalized B-value} = \frac{B - \langle B \rangle}{\sigma_B}$$

where B represents raw B-value and $\langle B \rangle$ and σ_B are the average and standard deviation of the C-alpha B-values for the particular PDB structure, respectively.

Performance measure

Formula 2:

We have calculated the Pearson correlation coefficient (CC) between the theoretical and experimental B-values to verify the performance of SVR method, as given by,

$$CC = \frac{\sum_{i=0}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=0}^m (x_i - \bar{x})^2][\sum_{i=0}^m (y_i - \bar{y})^2]}}$$

where x_i and y_i are the respective experimental and theoretical values of the normalized B-value of the i^{th} C-alpha atom, and \bar{x} and \bar{y} are their corresponding sample means.

Table 1: Total Heterotrophic and hydrocarbon degrading bacterial count

Window size	γ	C	CC
7	0.00098	1.5	0.42
9	0.00098	1.5	0.42
11	0.00069	2	0.43
13	0.00049	2	0.43
15	0.00069	1.5	0.47
17	0.00049	2	0.46
19	0.00069	1.5	0.45
21	0.00049	2	0.45

*CC is Pearson correlation coefficient