PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

On the relation between residue flexibility and local solvent accessibility in proteins

Hua Zhang,[1,2*] Tuo Zhang,[1,2] Ke Chen,[2] Shiyi Shen,[1,3] Jishou Ruan,[1,3] and Lukasz Kurgan[2*]

[1] College of Mathematical Science and LPMC, Nankai University, Tianjin, People's Republic of China
[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
[3] Chern Institute of Mathematics, Nankai University, Tianjin, People's Republic of China

ABSTRACT

We investigate the relationship between the flexibility, expressed with B-factor, and the relative solvent accessibility (RSA) in the context of local, with respect to the sequence, neighborhood and related concepts such as residue depth. We observe that the flexibility of a given residue is strongly influenced by the solvent accessibility of the adjacent neighbors. The mean normalized B-factor of the exposed residues with two buried neighbors is smaller than that of the buried residues with two exposed neighbors. Inclusion of RSA of the neighboring residues (local RSA) significantly increases correlation with the B-factor. Correlation between the local RSA and B-factor is shown to be stronger than the correlation that considers local distance- or volume-based residue depth. We also found that the correlation coefficients between B-factor and RSA for the 20 amino acids, called flexibility-exposure correlation index, are strongly correlated with the stability scale that characterizes the average contributions of each amino acid to the folding stability. Our results reveal that the predicted RSA could be used to distinguish between the disordered and ordered residues and that the inclusion of local predicted RSA values helps providing a better contrast between these two types of residues. Prediction models developed based on local actual RSA and local predicted RSA show similar or better results in the context of B-factor and disorder predictions when compared with several existing approaches. We validate our models using three case studies, which show that this work provides useful clues for deciphering the structure–flexibility–function relation.

Proteins 2009; 76:617–636.

Key words: B-factor; disordered regions; ordered regions; flexible region; residue depth; secondary structure; active site.

INTRODUCTION

Proteins undergo constant thermal fluctuations and other types of motions that span between rapid (picoseconds) vibration and relatively slow (microseconds to seconds) movements.[1] The structural flexibility associated with these motions allows implementation of various biological processes such as molecular recognition, enzyme catalysis, allosteric regulation, antigen–antibody interactions, and protein–DNA binding.[2–6] Experimentally available structural data that were derived based on X-ray crystallographic studies provide information on the atomic mobility, which is represented by the atomic displacement parameter, also known as the Debye–Waller temperature factor or B-factor. This parameter reflects the degree of dispersal of atomic electron density around the equilibrium position due to thermal motion and positional disorder. The B-factors have been studied from a variety of viewpoints including the relation between mobility and thermal stability,[7,8] in the context of applications in the prediction of active sites and binding sites,[9–12] in the design of potential function,[13] and in protein function analysis/discovery.[2,6,14–16] Molecular dynamic (MD) simulation is one of the most powerful computational methods used to describe and analyze protein flexibility. The main drawback of MD simulations is their high computational cost.[17–19] Several prediction methods that address protein flexibility and that investigate its relation with protein function were developed to overcome this limitation. They include structure-based[20–24] and sequence-based[25–30] methods, where in both cases B-factor was used as the enabling concept. Recent studies show that the structure-based methods, such as the Gaussian network model (GNM),[21] the mean-field-like model,[17] the elastic network model (ENM),[19] the protein fixed-point (PFP) model,[23] and the weighted contact number (WCN) model,[24] could provide better insights to the structure–dynamics–function relation-

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: NSERC (Canada), National Education Committee of China, NSFC; Grant number: 10671100; Grant sponsors: Liuhui Center for applied mathematics, the joint program of Tianjin and Nankai Universities, Alberta Ingenuity Fund, iCORE.
*Correspondence to: Hua Zhang, College of Mathematical Science, Nankai University, Tianjin 300071, People's Republic of China. E-mail: zerohua@gmail.com or Lukasz Kurgan, Department of Electrical and Computer Engineering, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, Canada T6G 2V4. E-mail: lkurgan@ece.ualberta.ca
Received 3 October 2008; Revised 5 December 2008; Accepted 16 December 2008
Published online 20 January 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22375

ship of proteins than the conventional MD simulations due to their ability to sample a wider range of collective motions.[31] However, such methods require the knowledge of the atomic coordinates. In contrast, the sequence-based predictors use only the protein sequences as the input and thus they are suitable for the analysis of the chains with the unknown structure.

Since the dynamic processes concerning folding and interactions with ligands are complex, a few different definitions of flexible regions were proposed. As an extreme manifestation of flexibility, a class of "natively unstructured" or "intrinsically disordered" regions was defined as the regions that are invisible in X-ray diffraction electron density maps. In the past decade, protein disorder has received considerable attention due to its important role in various protein functions.[32,33] A recent result by Jones group shows that 30–60% of all eukaryotic proteins may contain disordered regions.[34] To this end, a number of flexible region predictors have been developed[35,36] and some of them show strong correlations with B-factor values.[28,37] However, the conceptual connection between flexible and natively unstructured regions remains obscure.[38] Recently, a disorder prediction method RONN[39] has been applied to analyze the flexibility of aromatic amino acids in cap-binding proteins[40] based on the observation that the flexibility indices[27] computed from B-factor values and the mean values of disorder probability for each type of amino acid are highly correlated. Abovementioned observations suggest that the two manifestations of flexibility, B-factors and disordered regions of proteins, are closely related. At the same time, the disorder is also closely related to protein function.[41] Since the B-factor values are not available for the disordered residues, we cannot directly examine the relationship between B-factors and the disordered regions. Instead, we investigate whether a model for prediction of B-factor values developed based on the remaining (ordered) residues could be used to detect the disordered regions.

The solvent-accessible surface area (ASA) has been widely studied due to the fact that surface residues are directly involved in the interaction with other biological molecules.[42,43] The ASA was used in the context of protein function, stability, and fold recognition.[44–47] Several methods were developed for the prediction of relative solvent accessibility (RSA),[48–51] which is defined by the ASA of a residue in the protein divided by ASA observed in an extended conformation (Gly-X-Gly or Ala-X-Ala).

The relations between flexibility and a few related concepts such as contact density and contact number have been previously discussed.[22,24] The flexibility of a residue is also known to be correlated with its ASA.[52,53] Mobile sections of a protein often have high solvent accessibility and only a few scaffolding hydrogen bonds between the domains.[54] A sequence-based flexibility prediction method by Schlessinger and Rost[30] uses predicted binary RSA, which annotates a given residue as exposed or buried using a cutoff threshold, as its inputs, which provides further evidence of this relation. However, a detailed analysis of the relationship between the B-factor and the solvent accessibility was never attempted. As observed by Halle,[22] this relation cannot be accurately described using a simple linear function, that is, $B_i = a \times \mathrm{ASA}_i + b_0$ where i represents the ith residue in a protein sequence, since such model would result in B-factor values of all buried residues be the same and equal to b_0. Additionally, other residue descriptors such as distance- or volume-based residue depth indices, which are complementary of RSA and which allow describing the interior of the proteins,[55–57] could be also considered in the context of their relation with the flexibility. These descriptors were shown to be useful for the analysis of amide hydrogen/deuterium exchange rates in nuclear magnetic resonance (NMR) experiments,[58] for the analysis of the local packing arrangements in the protein core,[59] and for protein fold recognition.[47] At the same time, their relation with the flexibility has not been studied.

In this study we focus on the relation between the residue flexibility measured with the B-factor and the solvent accessibility. This relation is investigated in the context of different types of amino acids and secondary structures as well as using tripeptide-based exposure patterns. Since the motions in protein are not constrained to individual residues but they also involve neighboring residues creating a dynamic network,[4,59] we study the impact of the solvent accessibility of immediate and further neighbors of the investigated residue. To do that, we use least square linear regression model and we optimize the size of a local, with respect to the sequence, window using a large dataset of 972 chains from Refs. 23 and 24. We also contrast the relation between the flexibility and solvent accessibility with the relation between the flexibility and residue depth.

In our work we use the actual and the predicted RSA values. The former RSA values are computed from known protein structure, while the latter are predicted using the protein sequence. This allows applying our conclusions in the context of sequence-based prediction of the disordered regions. Using a new blind dataset, we use the predicted RSA values and our linear regression model to predict B-factor values, which in turn are used to find the disordered regions. We emphasize that this is accomplished in spite of the fact that these regions by default have no actual B-factor values. Finally, we apply the linear regression models that use either the actual or the predicted RSA values on three case studies which involve analysis of *Escherichia coli* RNase HI,[38] human interleukin-2[16] and human cyclin-dependant kinase-2 (CDK2)[15] proteins. The goal of these case studies is to show that the relation between solvent accessibility and B-factor values, which is quantified with the regression

Table I
Relationship Between Mean B′-Factor and Mean RSA Values for the 20 Amino Acids

AA type	No. of residues	Mean B′-factor	CC between B′-factor and RSA	Mean actual RSA	Stability scale (kcal/mol)
K	15650	0.345	0.446	0.472	2.12
E	17844	0.319	0.513	0.470	1.89
D	16823	0.240	0.509	0.426	1.75
P	13553	0.173	0.483	0.320	2.09
S	16945	0.137	0.559	0.313	1.66
Q	11007	0.131	0.495	0.388	2.16
N	12966	0.114	0.528	0.388	1.85
G	21760	0.098	0.542	0.287	1.17
R	13726	0.009	0.452	0.355	2.71
T	16117	−0.010	0.524	0.281	2.18
A	22977	−0.075	0.515	0.204	2.18
M	6467	−0.081	0.565	0.141	3.63
H	6964	−0.109	0.542	0.268	2.51
L	24375	−0.188	0.398	0.119	4.71
C	4136	−0.224	0.421	0.095	3.89
V	19507	−0.230	0.438	0.123	3.77
I	15777	−0.247	0.411	0.103	4.5
Y	10760	−0.286	0.358	0.177	5.01
F	11700	−0.289	0.347	0.117	5.88
W	4844	−0.298	0.355	0.142	6.46

The rows are in the descending order of the mean B′-factor values. The computations are based on the PDB972 dataset.

model, can be used to find rigid/flexible regions that in turn give useful clues in the context of protein function.

MATERIALS AND METHODS

Datasets

We use a dataset which was proposed in Refs. 23 and 24 and which was selected using PDB-REPRDB.[60] This set, referred to as PDB972, includes 972 protein chains of length ≥60 which are characterized by pairwise sequence identity ≤25% and which include structures that are solved by X-ray crystallography with resolution ≤2.0 Å and R-factors ≤0.2. The second dataset, referred to as PDB766, was introduced in Ref. 29 and contains 766 protein chains selected using the same criteria as the PDB972 dataset.

We also prepared a new dataset based on sequences that were deposited to Protein Data Bank (PDB)[61] between January 2007 and April 2008, and which were filtered to have low identity with the sequences in the PDB972 dataset and the sequences deposited to PDB before 2007. More specifically, the sequences deposited before 2007 and after 2007 were separately filtered using CD-hit program[62] with 95% identity threshold. The resulting sets concerning these two time periods are referred to as PDB95-B07 and PDB95-A07, respectively. Since the minimal identity threshold of otherwise highly efficient CD-hit equals 40%, we used NCBI's BLASTCLUST[63] to the union of PDB95-A07, PDB95-B07, and PDB972 with the local identity threshold set at 25% and default minimal length coverage of 90% (-S 25 -L 0.9 options). The new dataset was constructed by selecting one chain of length ≥60 with best resolution ≤2.0 Å and R-factors ≤0.2 from each of the clusters that contained no sequences from the PDB95-B07 and PDB972 datasets. This set, called PDB328, includes 328 chains that, as a result, have local 25% identity with each other and also with the PDB95-B07 and PDB972 datasets. The PDB identifiers of chains from the PDB328 dataset are given in the supporting information Table I.

The PDB972 dataset is used to study the relation between the solvent accessibility and the flexibility. The PDB766 and PDB972 datasets are used to contrast the proposed RSA-based linear model for prediction of B-factor values with other sequence based methods for B-factor prediction. The PDB328 dataset is used to investigate the relation between B-factor, disordered regions and predicted RSA. This dataset is used to compare disorder region prediction obtained based on the findings in this paper with results of a recent disorder region prediction method. The low identity with respect to the PDB95-B07 and PDB972 datasets allows for an unbiased (with respect to sequences used to develop the prediction methods) comparison.

B-factor and disordered regions

Experimental B-factor of an atom is defined as $8\pi^2\langle u^2\rangle$ using the isotropic mean square displacement, u^2, averaged over the lattice.[30] Since B-factor values depend on the experimental resolution, crystal contacts, and the refinement procedures, they have to be normalized to allow comparisons between different structures. Following Refs. 8 and 30, B-factors of C_α atoms for each chain were extracted from PDB files and normalized using:

$$B' = \frac{B - \overline{B}}{\sigma} \tag{1}$$

where B is the actual B-factor, $\bar{B}$ is the average B-factor in a given chain, and σ is the standard deviation of B-factors for all C_α atoms in a given chain. This normalization was applied to B-factors in the PDB972 and PDB766 datasets. The disordered regions in the PDB328 dataset were identified as the residues that do not have coordinates in X-ray structures according to the 'REMARK465' record in the header of the corresponding PDB entry.

Solvent exposure properties

The actual ASA values in the three datasets were computed with DSSP program,[64] which also assigns eight-state secondary structures to each residue. Predicted ASA values were derived using Real-SPINE[65] which is motivated by high quality of predictions generated by this method, that is, the authors reported correlation coefficient of 0.74 and mean absolute error of 0.142 between the predicted and the actual ASA. Following Ref 65, RSA was computed by the ASA of a residue normalized by the ASA of this residue in its extended tripeptide (Ala-X-Ala) conformation.[49] The predictions with Real-SPINE on the PDB972 and PDB766 dataset yielded 0.71 and 0.72 correlation coefficient, and 0.145 and 0.143 mean absolute error, respectively, and thus we assume that this method did not overfit the two datasets.

The distance-based depth is defined as the minimum distance between an atom and a dot of solvent accessible surface[55] or its closest solvent accessible neighbor.[56] The residue depth ($\mathrm{RD_{dis}}$) is the average atom depth of all atoms composing a given residue. Similarly as in Refs. 66 and 67, the MSMS program[68] was first executed with a probe radius of 1.4 Å to obtain a list of vertices that represent the protein surface. The atom depth, that is, the distance between an atom and its nearest vertex, was calculated, and the average atom depth of all atoms except the hydrogen atoms for a given residue was assumed as its depth.

In the case of the volume-based depth,[57] given an atom i and a sampling radius r, a depth index $D_{i,r}$ is defined as $D_{i,r} = 2V_{i,r}/V_{0,r}$, where $V_{i,r}$ is the exposed volume of a sphere of radius r centered on atom i and V_0,r is the exposed volume of the same sphere when centered on an isolated atom. Following Refs. 57 and 67, we computed the residue depth values ($\mathrm{RD_{vol}}$) as the depth of C_α atoms with a sampling radius of 9 Å using SADIC program.

Linear regression models

We use a linear regression model over a local window in the protein sequence to express the relation between the solvent exposure, expressed using RSA values, and flexibility, expressed using normalized B-factor values. The flexibility of the central residue in the window, denoted as B′-factor (normalized B-factor), is defined as:

$$\hat{B}'_i = \sum_{k=-h}^{h} w_k \cdot RSA_{i+k} + b \tag{2}$$

where b is the intercept and $\hat{B}'_i$ represents the estimated (predicted) B′-factor of the central residue i using RSA values in the window of size $h = 0, 1, 2, \ldots$, (the window includes $2h + 1$ residues), and where weighs w_k are determined using the least squares fit between the estimated (predicted) B′-factor and the actual B′-factor values. In our study, RSA_i correspond to either the actual RSA values derived with DSSP (denoted by $DsspRSA_i$) or the predicted values (denoted by $PredRSA_i$). The models that use the actual and the predicted RSA and which correspond to different window sizes are referred to as DsspRSA*s* and PredRSA*s*, where $s = 2h + 1$, respectively. In order to evaluate the ability of these models to generalize to new data, we performed fivefold cross validation tests by following Ref. 29. The computation of the weights was performed using Weka.[69]

Evaluation measures

Based on Refs. 24 and 29, the Pearson correlation coefficient (CC), which is defined as

$$CC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}} \tag{3}$$

where x_i is the observed B′-factor and y_i is the RSA value or predicted B′-factor for the ith residue in the sequence, is used to quantitatively measure the relationship between B′-factor and solvent accessibility and to evaluate the quality of the proposed linear models. If CC is close to 1, then $\{x_i\}$ and $\{y_i\}$ are fully correlated. If CC is close to 0 then the two variables are not correlated, and in the case when CC is close to −1 then the variables are anticorrelated. The absolute CC values quantify the degree of the correlation.

We note that the correlation can be measured at the residue level[29] or at the chain level.[23,24] In the former case, all residues in a given dataset are merged together and one CC value is computed. In the latter case, CC is computed for each chain separately and next these values are averaged to compute the correlation over the entire dataset. In this article, we report CC at the residue level, unless stated otherwise, in which case we will use term average correlation coefficient (ACC) that refers to the CC at the chain level.

Statistical tests

We use statistical tests to verify whether the distributions used in the course of this paper are different. We used *t*-test to compare two normal distributions and Shapiro–Wilk test to verify whether a given distribution is normal. These statistical tests were performed with statistical package R.[70]

RESULTS

First, we investigate the relation between the flexibility and solvent accessibility for individual residues. This is followed by an analysis of this relation in a context of neighboring residues, including both immediate neighbors and a local window in the protein sequence. We also compare our window-based model that can be used to predict B′-factor values from the solvent accessibility (both actual accessibility and accessibility predicted from the protein sequence) with relevant methods for prediction of B′-factor values. Next, we investigate the relation between residue depth and B′-factor and contrast it with the relation between solvent accessibility and B′-factor. Finally, we explore the relation between the solvent accessibility predicted from the protein sequence and B′-factor in the context of the application into prediction of disordered regions. This section is concluded with an application into three case studies.

Relation between B′-factor and RSA at the single residue level

The CC and ACC between B′-factor and solvent accessibility for residues in the PDB972 dataset equal 0.47 and 0.48 for the actual ASA and 0.51 and 0.52 for the actual RSA, respectively. Since the RSA values have higher correlation with B-factor than the ASA values, only RSA values are used to quantify the relationship between B-factor and the solvent accessibility.

Relation between B′-factor and RSA at the single residue level for different amino acid types

We divided the 283,898 residues in the PDB972 dataset into 20 subsets according to the type of the amino acids and examined the correlations between B′-factor and mean RSA for each of these subsets. Table I lists the mean B′-factor and mean RSA values for the 20 amino acids together with the CC values between B′-factor and RSA. The CC between the 20 mean B′-factor values and 20 mean RSA values over the standard amino acids equals 0.93, which suggests that higher mean B′-factor implies larger mean RSA, that is, flexible residues are usually on the protein surface so that a sufficient amount of solvent accessibility allows them to carry out specific functions. For instance, the table shows that the charged residues, which include Lys (K), Glu (E), and Asp (D), are not only the most flexible but also the most exposed. The CC values between B′-factor and RSA for the amino acids yield a new amino acid index which we call flexibility-exposure correlation index (FECI). This index reflects the strength of the relation between B′-factor and RSA. We observe that FECI is not strongly correlated with neither mean B′-factor, CC = 0.62, nor mean RSA, CC = 0.51. This implies that the strength of this relation is mediated by some other property. We searched all amino acid indices deposited in the AAindex database,[71] and found that FECI is strongly correlated with the stability scale,[72] see Table I. The CC between the stability scale and the FECI index equals −0.85 and the corresponding linear regression line is shown in Figure 1. The stability scale, which was derived from the knowledge-based atom–atom potentials, characterizes the average contributions of individual residues to the folding stability. The negative correlation suggests that the individual residues with higher (lower) FECI in general have lower (higher) stability scale value, that is, they have lower (higher) average contributions to the folding stability, and vice versa. For example, Trp (W) has the lowest mean B′-factor, the second lowest CC between B-factor and RSA, and the largest value according to the stability scale, but it does not have the lowest mean RSA. The lowest mean B′-factor together with low CC between B-factor and

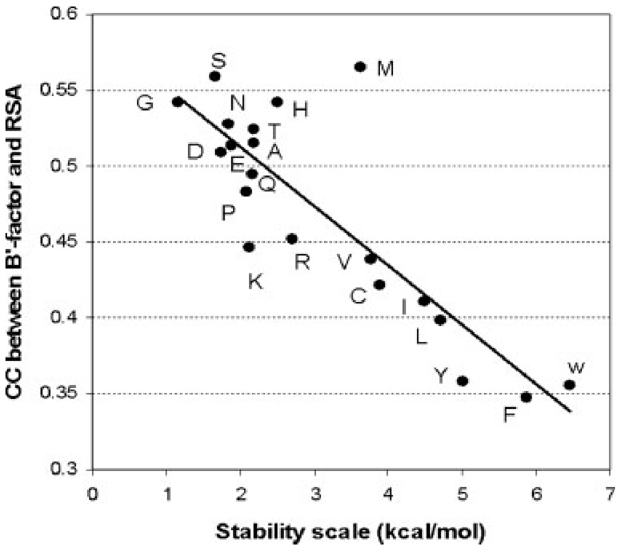

Figure 1

The relation between flexibility-exposure correlation index (*y*-axis) and the stability scale (*x*-axis) for the 20 amino acids. Correlation coefficient corresponding to the shown linear regression line equals −0.85. The computations are based on the PDB972 dataset.

Table II
Relationship Between Mean B′-Factor and Mean RSA for Eight-State and Three-State Secondary Structures

SS type	No. of residues	Mean B′-factor	CC between B′-factor and RSA	Mean actual RSA
Eight-state secondary structures				
T	33918	0.354	0.499	0.452
—	57834	0.270	0.520	0.296
S	25506	0.266	0.481	0.350
G	12128	0.151	0.450	0.335
B	3748	−0.147	0.446	0.183
H	88672	−0.160	0.435	0.236
I	64	−0.230	0.620	0.117
E	62028	−0.347	0.460	0.148
Three-state secondary structures				
C	117258	0.294	0.497	0.353
H	100864	−0.123	0.446	0.248
E	65776	−0.336	0.460	0.150

The rows are in the descending order of the mean B′-factor values. The computations are based on the PDB972 dataset.

RSA imply that surface Trp residues tend to be relatively rigid, when compared to other residues. Recent studies show that surface Trp residues strongly contribute to the folding stability.[73,74] More specifically, they show that single mutation (W62G) of hen egg-white lysozyme results in a less stable structure than that of the wild-type,[74] which supports our finding.

Relation between B′-factor and RSA at the single residue level for different secondary structures

DSSP program[64] was used to assign the eight-state secondary structures, which include α-helix (H), 3_{10}-helix (G), π-helix (I), β-sheet (E), β-bridge (B), hydrogen bonded turn (T), bend (S), and random coil or loop (-), for all residues in the PDB972dataset. We reduced the eight-state structures into three states by grouping H, G, and I as helix (H), B and E as strand (E), T, S, and - as coil (C); the same conversion is performed in the EVA server.[75]

We separated all the residues in the PDB972 dataset based on their eight-state and three-state structures into eight and three subsets, respectively. The corresponding mean B′-factor and mean RSA values, as well as the CC between B′-factor and RSA for the residues in each of the eight and three secondary structure states are given in Table II. The hydrogen bonded turn (T) is characterized by the highest flexibility, that is, the mean B′-factor equals 0.354, and the highest mean RSA that equals 0.452. The random coil (-) and bend (S) have almost the same mean B′-factor values, that is, the corresponding *P*-value using two-sided *t*-test equals 0.6259. Among the three helix types (H, G, I), 3_{10} helix is the most flexible, α-helix is more rigid, and the π-helix is the most rigid. These differences are likely due to the geometry of the helices, in which 3, 3.6, and 4.1 residues per turn and translation of 2 Å, 1.5 Å, and 1.15 Å for the 3_{10}, α-, and π-helices imply increasingly tighter packing. We note that the number of π-helices in our dataset is relatively small, thus potentially limiting reliability of our observation. Among the two strand types, β-sheet is less flexible than β-bridge. We observe that the mean B′-factor of β-bridges is close to that of α-helix. When examining the three-state secondary structures, the table shows, as expected, that coils are the most flexible, while strands are the most rigid. The order of mean B′-factor values is in consistent with the order of mean RSA values for the three-class secondary structures. The distributions of B′-factor and RSA values in three-state secondary structures are shown in Figure 2. We note that the CC values between B′-factor and RSA are relatively similar across all secondary structure types, which suggests that the relation between these two structural descriptors does not vary with the secondary structures. Finally, the mean RSA values show that strands tend to be more buried than helices and coils, while the most solvent-exposed secondary structure is the bonded turn.

The impact of RSA of the neighboring residues on the residue flexibility

The distributions of B′-factor values for residues binned according to RSA values in the PDB972 dataset, see Figure 3, indicate that the exposed residues are on average more flexible than the buried residues. The Figure also shows that certain buried residues could be more flexible than some exposed residues, which comes from the overlap between the corresponding distributions. This is likely due to the fact that a protein is a dynamic network formed by the connected residues, which means that local packing and arrangement of neighboring residues, especially the exposure or burial of the adjacent neighbors, strongly impact the flexibility of a given residue. To this end, we investigate the exposure patterns of tripeptides based on the RSA cutoffs. To avoid using an arbitrary threshold we use twenty RSA cutoffs ranging between 0 and 95% with a step of 5%. A given residue is defined as exposed (*e*) if its RSA is larger than the cutoff value, and otherwise it is defined as buried (*b*). A given (central) residue that assumes the exposure state *x* (one of two states, *e* and *b*) may have two buried (*bxb*), two exposed (*exe*), or one buried and one exposed (*exb* and *bxe*) adjacent neighbors. The mean B′-factor values of the central residues for each of the six possible tripeptide exposure patterns for each RSA cutoff are plotted in Figure 4. The Figure shows a consistent increase of mean B′-factors with the increasing values of the RSA cutoff. We observe that irrespective of the exposure state of the central residue, the exposure to the solvent of the adjacent residues promotes flexibility of the central residue, while the burial of the adjacent residues inhibits the flexibility of the central residue. Contrary to the intuition, given any RSA cutoff value, the exposed residues with two bur-

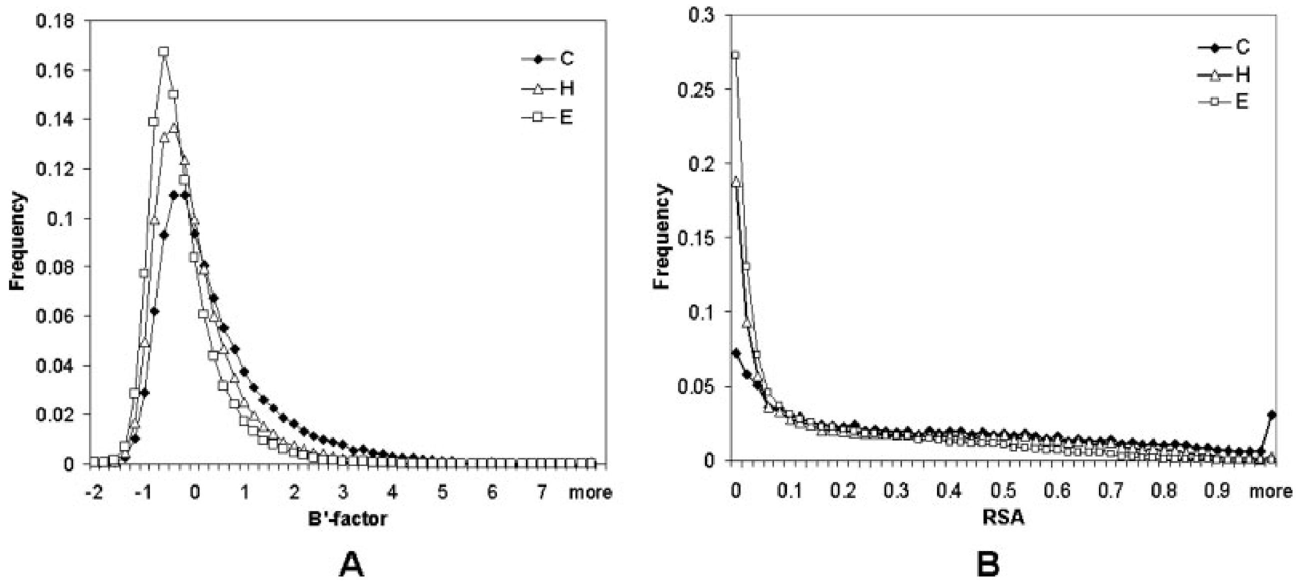

Figure 2

Distributions of B′-factor values (Panel **A**) and RSA values (Panel **B**) for the three major types of the secondary structures. The computations are based on the PDB972 dataset.

ied adjacent neighbors (i.e., *beb* pattern) have lower mean B′-factor values than the buried residue with two exposed adjacent neighbors (i.e., *ebe* pattern). This implies that the two buried neighbors strongly influence the flexibility of the central residue making it more rigid than the buried residue which is flanked by two exposed residues.

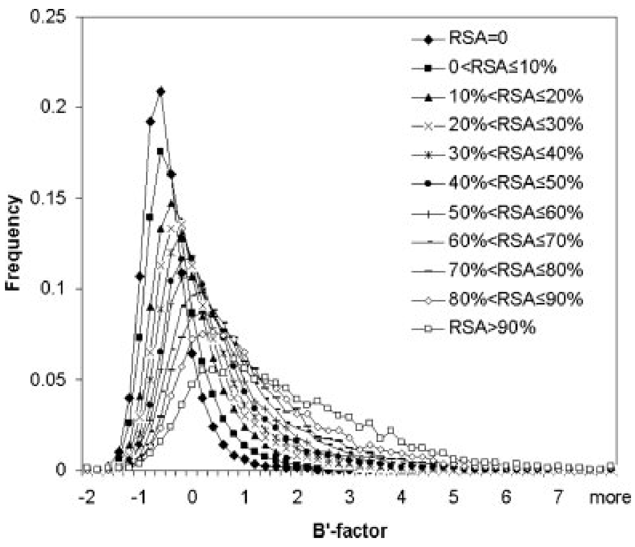

Figure 3

Distributions of B′-factor values for residues binned according to their RSA values. The B′-factor values were discretized into 0.2 wide intervals. The computations are based on the PDB972 dataset.

As an example, we further investigate two cases with RSA cutoffs equal to 20 and 25%. The latter threshold is applied in two-class (buried/exposed) RSA prediction,[65] while the former cutoff value results in the most balanced division into exposed and buried residues. The distributions of B′-factor values for each of the six tripeptide exposure patterns and the corresponding mean B′-factor values are shown in Figure 5 and Table III, respectively. We again observe that for both the cutoffs the residues with *beb* pattern have smaller mean B′-factor values than the residues within *ebe* pattern. We performed two-sided *t*-test that compares the corresponding two distri-

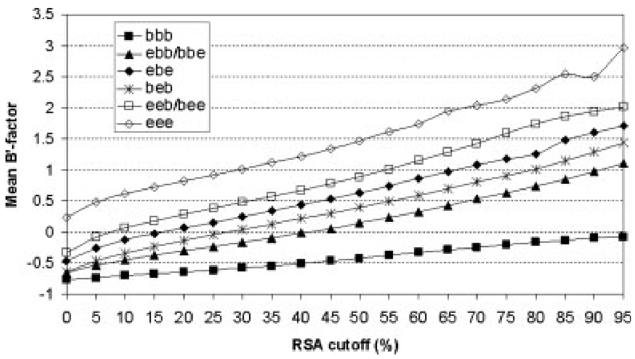

Figure 4

Mean B′-factor values of central residues for the six tripeptide exposure patterns (*y*-axis) which are defined based on different RSA cutoffs (*x*-axis). The computations are based on the PDB972 dataset.

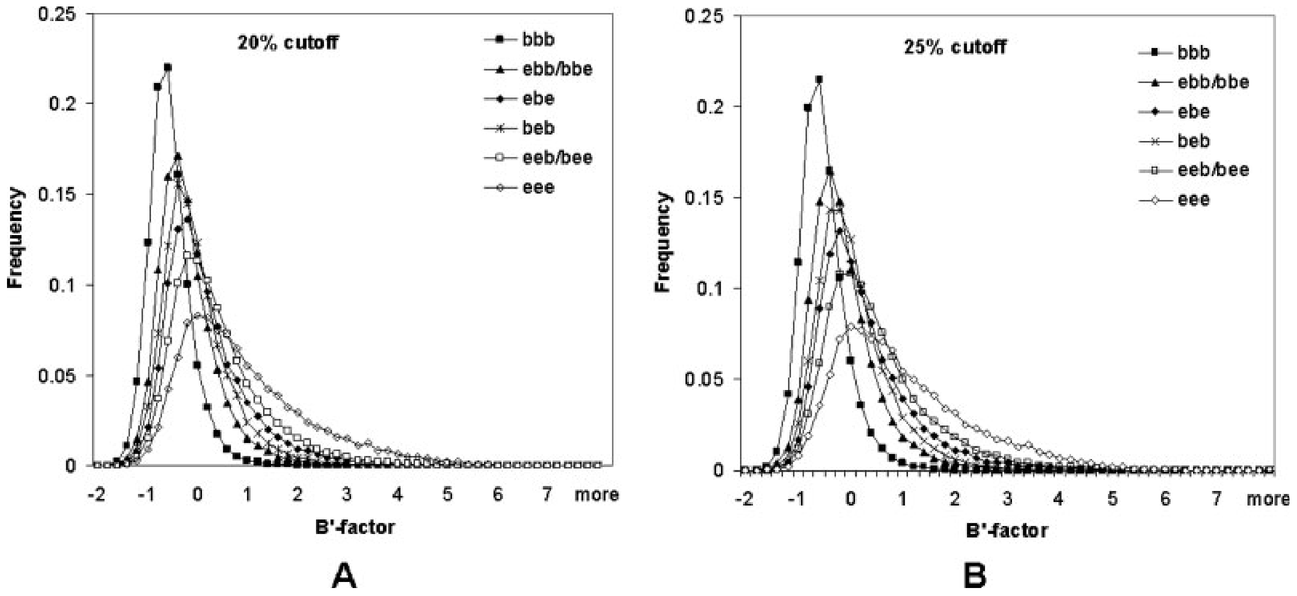

Figure 5

The distributions of B′-factor values of central residues for the six tripeptide exposure patterns and two RSA cutoffs. Panel **A** shows results for RSA cutoff equal 20%. Panel **B** shows results for RSA cutoff equal 25%. The computations are based on the PDB972 dataset.

butions of B′-factor values for both RSA thresholds, which shows that they are different with *P*-values <2.2e-16. The same tests performed for all other pairs of distributions also shows that they are different with *P*-values <2.2e-16. The above results suggest that the neighboring residues have significant impact on the flexibility of the central residue.

The impact of the local solvent accessibility on the residue flexibility

In addition to the impact from the adjacent residues discussed in the above section the solvent accessibilities of other residues that are further along the sequence (in a local sequence window) may also contribute the flexibility. We developed linear regression models that take RSA/ASA values of residues in a local window as an input to compute the B′-factor value of the central (in the window) residue. These models are used to quantify the relation between local (with respect to the sequence) RSA/ASA and B-factor. We also use them to study the impact of the window sizes on the strength of this relation. The regression models are computed based on five-fold cross validation on the PDB972 dataset. Figure 6A shows the ACC values that quantify correlation between the outputs of the linear regression models and the B′-factor values for the window sizes between 1 and 25 for DsspRSA (actual RSA), DsspASA (actual ASA), PredRSA (RSA predicted using Real-SPINE) and PredASA (ASA predicted with Real-SPINE) models. The Figure shows that local RSA values have stronger correlation with B′-factors than the local ASA values. The differences when using the actual and the predicted solvent accessibility are consistent across different window sizes. The improvement in ACC due to the increased window size is larger when considering small sizes between 1 and 3,

Table III

The Comparison of Mean B′-Factor Values for the Six Tripeptide Exposure Patterns and Two RSA Cutoffs

Exposure of the central residue	Tripeptide exposure pattern	20% cutoff		25% cutoff	
		No. of residues	Mean B′-factor	No. of residues	Mean B′-factor
Buried	bbb	67576	−0.646	79789	−0.613
	bbe/ebb	54223	−0.297	58158	−0.229
	ebe	27145	0.061	25064	0.150
Exposed	beb	21507	−0.134	23695	0.048
	bee/eeb	65211	0.280	60527	0.376
	eee	46292	0.813	34721	0.917

The computations are based on the PDB972 dataset.

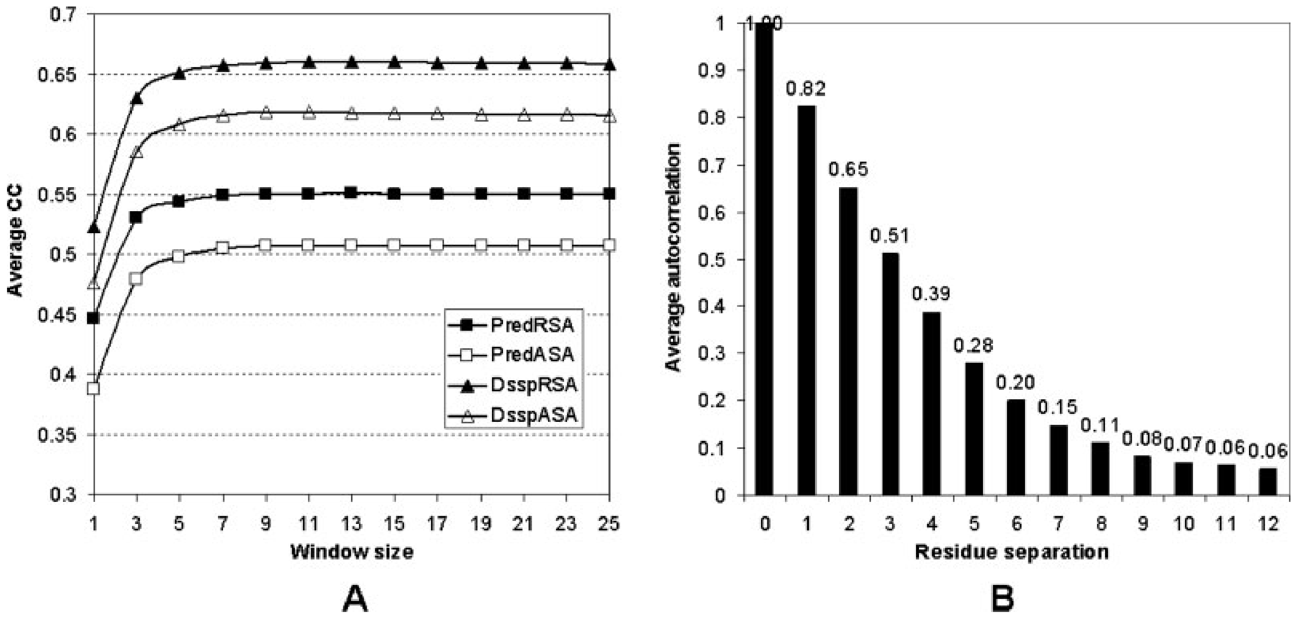

Figure 6

Panel (**A**) Strength of the relation between B′-factor and local solvent exposure which is measured with ACC (*y*-axis) and which is computed using varying window sizes (*x*-axis). The solvent exposure is modeled using the actual RSA (DsspRSA) and ASA (DsspASA) values, as well as predicted RSA (PredRSA) and predicted ASA (PredASA) values. The regression models are computed based on fivefold cross validation on the PDB972 dataset. Panel (**B**) The average autocorrelations of B′-factor values (*y*-axis) for the residue separations between 0 and 12 (*x*-axis). The corresponding average autocorrelation values for each residue separation were computed using formula from Ref. 27 based on the PDB972 dataset; the values are shown above the bars.

that is, ACC values increase from 0.52 to 0.63 for DsspRSA and from 0.45 to 0.53 for PredRSA. In contrast, for window sizes of 9 and above, there is virtually no improvement in the ACC. The maximal ACC equals 0.66 for DsspRSA, 0.62 for DsspASA, 0.55 for PredRSA, and 0.51 for PredASA. The main reasons for the lack of improvement when using large window sizes are the relatively large separation between the residues (and thus potential lack of interactions) and the decrease of average autocorrelations of B′-factors. As shown in Figure 6(B), when the residue separation is ≥5 (i.e., the window size ≥11), the corresponding average autocorrelations are lower than 0.3.

We also directly compare the CC values which are derived using the linear regression models with the window sizes of 9 and 1 for each chain in the PDB972 dataset, see Figure 7. In the case of using the actual RSA values (DsspRSA model), see Figure 7(A), significant majority of chains show improvement in the correlation due to the usage of the local RSA values, that is, 960 out of 972 protein chains are located above the diagonal which denotes points where the correlation is the same for both window sizes. Similarly, when using the predicted RSA values (PredRSA model), see Figure 7(B), the improvement is observed for 920 out of 972 chains. Similar comparison was also performed in the case of individual AAs and secondary structures, and overall the conclusions are that the local RSA values contribute to the increase of CC irrespective of the type of the amino acid and the type of the secondary structure (data not shown).

Based on the above observations, the window size of 9 is selected to model the relation between the residue flexibility and the local RSA. We use two models, DsspRSA9 and PredRSA9, which are based on the actual local RSA and the predicted local RSA, respectively.

Comparison between DsspRSA and PredRSA models and other existing methods for prediction of B′-factor

As shown above, local RSA is strongly correlated with the B′-factor, which could be exploited to build a simple model for the prediction of residue flexibility. Table IV summarizes the prediction quality, measured based on the ACC and CC between the actual and the predicted B′-factor values, of several existing method for B′-factor prediction. We include two methods that predict the B′-factor values from the protein structure, WCN[24] and GNM,[21] which are compared against our regression model that is based on the actual DSSP values. We also report results for three methods that predict the B′-factor values from the protein sequence, a support vector regression (SVR) method by Yuan *et al.*,[29] a neural network method PROFbval,[30] and RONN method[39] which is primarily used to predict protein disorder. The latter three approaches are compared against our linear model

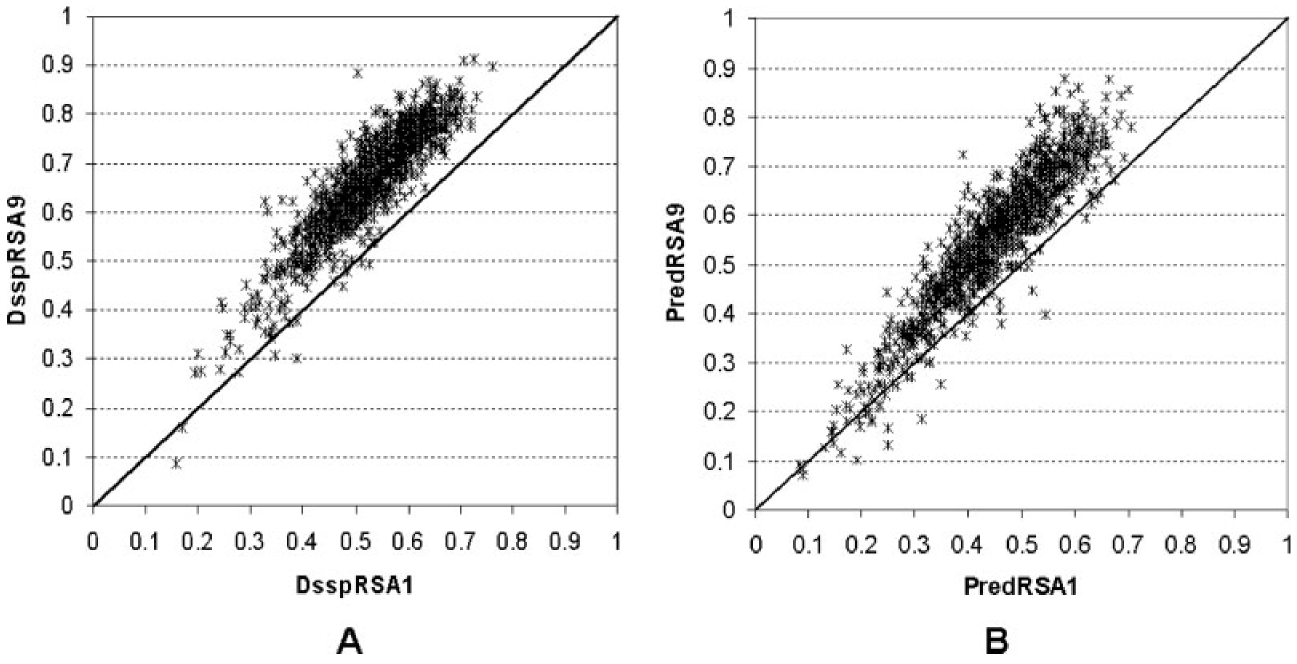

Figure 7

Comparison of CC between RSA and B′-factor for individual chains in the PDB972 dataset. Panel (**A**) compares the CC values when using the actual RSA of an individual residue (*x*-axis) and when using the actual RSA values in a window of size 9 (*y*-axis). Panel (**B**) compares the CC values when using RSA values predicted with Real-SPINE for a single residue (*x*-axis) and when using the predicted RSA values in a window of size 9 (*y*-axis).

that uses the predicted RSA as the input. RONN was used to verify whether a specific disorder predictor could model the flexibility measured with the B-factor. The structure-based methods, WCN and GNM, yield ACC values of 0.61 and 0.56, which were reported in Ref. 24, on the PDB972 dataset, while our DsspRSA9 model achieves ACC of 0.66. In the case of sequence-based methods, our PredRSA9 model provides the best result on the PDB766 dataset, that is, ACC = 0.56 and CC = 0.55, when compared with the SVR (CC = 0.53), the PROFbval (ACC = CC = 0.50 due to the normalization of the network outputs), and RONN (ACC = 0.14). We observe that probabilities outputted by RONN, which are designed for the prediction of the disorder, show weak correlation with the B′-factor.

Analysis of the linear regression models for the relation between local RSA and B′-factor

The linear model which represents the relationship between the B′-factor and the actual RSA values in window of size 9 (DsspRSA9) that was computed using the entire PDB972 dataset follows

Table IV

Comparison of Prediction Quality

	Structure based methods				Sequence based methods					
	DsspRSA9		WCN	GNM	PredRSA9		SVR	PROFbval		RONN
Datasets	ACC	CC	ACC	ACC	ACC	CC	CC	ACC	CC	ACC
PDB972	**0.66**	**0.63**	0.61	0.56	**0.55**	**0.53**	—	—	—	0.14
PDB766	**0.65**	**0.63**	—	—	**0.56**	**0.55**	0.53	0.50	0.50	0.14

Measured using CC and ACC between the predicted and the actual B′-factor values for the proposed linear regression models (DsspRSA9 and PredRSA9) and five existing method. The competing methods include two structure based algorithms, WCN[24] and GNM,[21] and three sequence based algorithms, SVR,[29] PROFbval,[30] and RONN.[39] The results of DsspRSA9 and PredRSA9 are based on fivefold cross validation on the PDB972 and PDB766 datasets; the remaining methods, which were designed in the corresponding references, were tested on the entire dataset. The results of DsspRSA9 and PredRSA9 are shown in bold, and the results denoted with "—" could not be computed due to the unavailability of the corresponding programs.

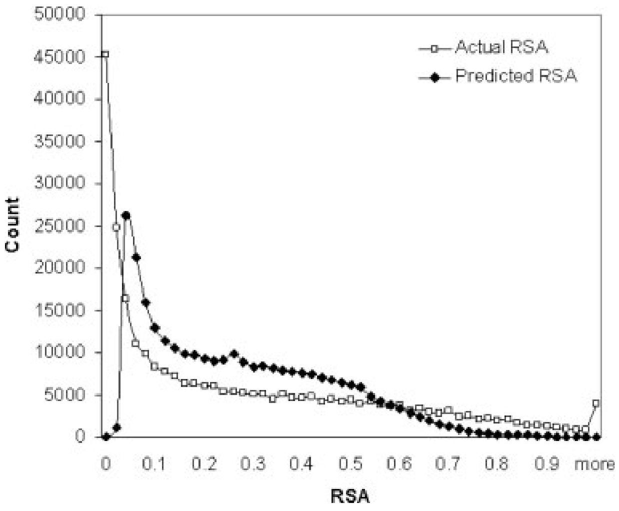

Figure 8

The distributions of the actual and the predicted RSA values computed from the PDB972 dataset. The RSA values were discretized with bin size of 0.02 (x-axis) and counted in intervals where RSA = 0, 0 < RSA ≤ 0.02, 0.02 < RSA ≤ 0.04, . . . , RSA > 1.0.

$$\hat{B}'_i = 0.0797\text{DsspRSA}_{i-4} + 0.1590\text{DsspRSA}_{i-3} + 0.3305\text{DsspRSA}_{i-2} + 0.6260\text{DsspRSA}_{i-1} + 1.1837\text{DsspRSA}_i + 0.6267\text{DsspRSA}_{i+1} + 0.3367\text{DsspRSA}_{i+2} + 0.1942\text{DsspRSA}_{i+3} + 0.1016\text{DsspRSA}_{i+4} - 0.9826$$

where i represents the ith residue in the protein sequence and $\hat{B}'_i$ denotes B′-factor estimate (prediction) for the ith residue.

The model that uses the predicted RSA values, PredRSA9, is shown below

$$\hat{B}'_i = 0.1430\text{PredRSA}_{i-4} + 0.2144\text{PredRSA}_{i-3} + 0.4184\text{PredRSA}_{i-2} + 0.7175\text{PredRSA}_{i-1} + 1.3322\text{PredRSA}_i + 0.7257\text{PredRSA}_{i+1} + 0.4141\text{PredRSA}_{i+2} + 0.2643\text{PredRSA}_{i+3} + 0.1607\text{PredRSA}_{i+4} - 1.1493$$

Both regression models show that the weight values decrease linearly, with a factor of 0.5, with the linear distance from the central, with respect to the window, residue. This decrease is symmetric, that is, towards both the N-terminus and C-terminus, and all weights are positive, which means that RSA values have promoting effect on the flexibility of the central residue. However, we observe that the weights in the PredRSA9 model are larger than the weights in the DsspRSA9 model. This is likely due to the differences in the distributions of the actual and the predicted RSA values, see Figure 8. The Figure shows that Real-SPINE, which was used to predict RSA values, tends to underpredict the highly exposed residues and to overpredict the deeply buried residues. In particular, residues in the PDB972 dataset that are fully buried, that is, their RSA equals zero, are predicted with a nonzero value by the prediction method. Similarly, majority of the residues that are significantly exposed, that is, RSA values >0.7, are predicted with lower ASA value, which results in lower values of the weights. However, the intersect values in the PredRSA9 and DsspRSA9 models, that is, −1.1493 and −0.9826, respectively, compensates for the differences in the weight values.

Relation between the B′-factor and the residue depth

The residue depth, which could be defined based on distance (RD_{dis}) and volume (RD_{vol}), is an alternative to solvent accessibility which allows for a better quantification for the residues in the interior of protein. We examine the relationship between representative residue depth indices and the B′-factor at the single residue level. This relation is measured based on the ACC and CC values between B′-factor, RSA, RD_{dis}, and RD_{vol}, see Table V. The RD_{vol} has the highest correlation with the B′-factor (ACC = 0.61 and CC = 0.59) when compared with RSA (ACC = 0.52 and CC = 0.51) and RD_{dis} (ACC = −0.39 and CC = −0.37). The reason for that is that the volume-based depth values, which are computed using spheres with sampling radius of 9 Å, include more information about the local residues when compared with RSA and distance-based depth. Relatively high correlation between RSA and RD_{vol} (ACC = 0.87 and CC = 0.87) shows that these two descriptors are related. The RD_{dis} is shown to have the lowest correlation with the B′-factor, which is likely due to the fact that it produces similar values for all surface residues (irrespective of their solvent exposure) and the fact that this depth index is dependent on the protein size.

In spite of the higher correlation at the single residue level between RD_{vol} and B′-factor when compared with the relation between RSA and B′-factor, usage of the local RSA results in improving the strength of the correlation that becomes higher than the correlation when considering local RD_{vol}, see Figure 9. The computations were performed based on fivefold cross validation on the PDB972 dataset applying the linear regression model over different window sizes. The Figure shows that ACC values for

Table V

Correlation Matrix That Lists ACC and CC Values Computed Between B′-Factor, RSA, RD_{dis}, and RD_{vol}

Descriptor		B′-factor	RSA	RD_{dis}	RD_{vol}
B′-factor	ACC	1.0	**0.52**	**−0.39**	**0.61**
	CC	1.0	**0.51**	**−0.37**	**0.59**
RSA	ACC		1.0	−0.59	0.87
	CC		1.0	−0.55	0.87
RD_{dis}	ACC			1.0	−0.66
	CC			1.0	−0.63
RD_{vol}	ACC				1.0
	CC				1.0

The computation was performed based on the PDB972 dataset. The bolded values concern ACC and CC values between B′-factor and RSA, RD_{dis}, and RD_{vol}.

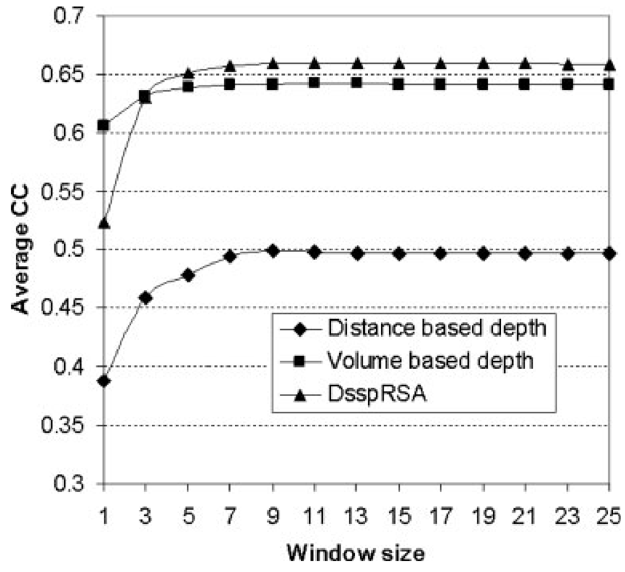

Figure 9

Strength of the relation between B′-factor and local RSA, local distance based depth (RD_{dis}), and local volume based depth (RD_{vol}), which is measured with ACC (*y*-axis) and which is computed using varying window sizes (*x*-axis). The regression models are computed based on fivefold cross validation on the PDB972 dataset.

all three descriptors display improvements with the increase of the window size of up to 9. The further increase of the window size has no effect on the correlation. Most importantly, RD_{vol} shows a relatively small increase of ACC from 0.61 to 0.64 due to added value of the local information, while usage of local DsspRSA brings the ACC to 0.66.

Relation between B-factor and predicted RSA in the context of disorder

We study the relation between the disordered regions in the protein sequence, which constitute an extreme manifestation of the flexibility, and the B′-factor values that were predicted using a linear regression model from the predicted RSA values, that is, B′-factor values predicted from the sequence. We use the PDB328 dataset, which incorporates 4470 disordered residues and 74,483 ordered residues. The motivation that supports such relation comes from the differences in the distributions of the predicted RSA values for the disordered and the ordered residues, see Figure 10(A). Figure 10(B) shows the distributions of the predicted B′-factor values for the disordered and ordered residues in the PDB328 dataset. These predictions are based on the PredRSA9 model that was built on the PDB972 dataset. In both the cases we observe that the distributions are significantly different, which should allow for building a well-performing predictor.

We use Receiver Operating Characteristic (ROC)[76] analysis to investigate whether predicted RSA values could be used to find disordered regions. We apply a series of thresholds on the outputs of a B′-factor predictor to classify the residues as ordered and disordered. The ROC curve shows the relation between TP rate (sensitivity) and FP rate (1-specificity) for each threshold, where

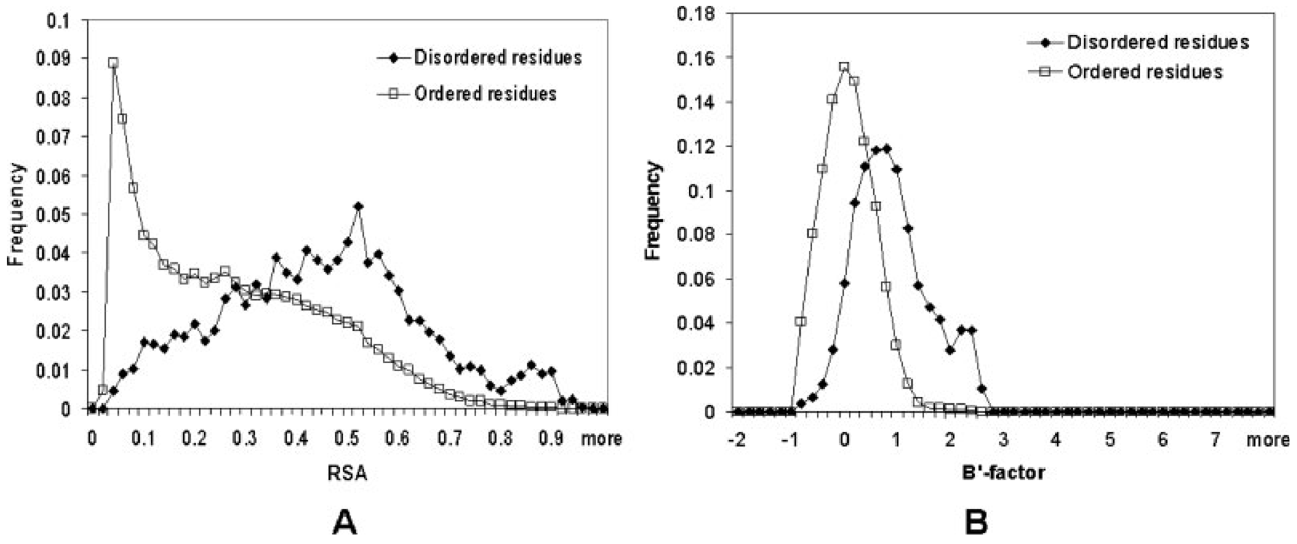

Figure 10

Panel (**A**) Distributions of RSA values for the disordered and ordered residues. The RSA values were discretized into 0.02 wide intervals. The computations are based on the PDB328 dataset. Panel (**B**) Distributions of the predicted B′-factor values for the disordered and ordered residues. The B′-factor values were discretized into 0.1 wide intervals. The prediction model was developed using the PDB972 dataset and the plot is based on the predictions performed on the PDB328 dataset.

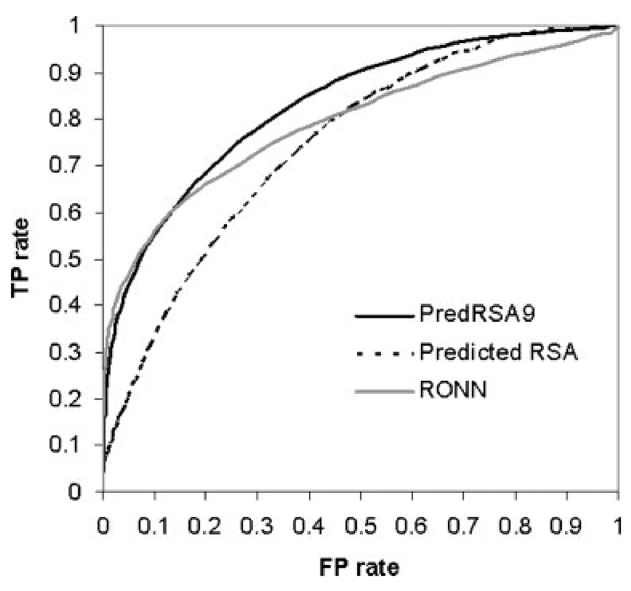

Figure 11

ROC curve for disorder region predictions on the PDB328 dataset. We compare the predictions of the PredRSA9 model, prediction when using RSA of an individual residue as the input, that is, PredRSA1 model, and the predictions generated with RONN.

the sensitivity is defined as the ratio between the number of correct predictions for disordered residues and the total number of the actual disordered residues, and the specificity is defined as the ratio between the number of correctly predicted ordered residues and the total number of the actual ordered residues. Figure 11 shows the ROC curves for the PredRSA9 model, which is based on the linear regression model over a window of size 9, the linear regression model that uses only the RSA values of an individual residue as the input, and the RONN method[39] that specializes in prediction of disordered regions. The areas under the ROC curve (AUC), which were calculated with ROCR[77] and which quantify the overall performance independently of the threshold values, equal 0.83 for PredRSA9, 0.75 for predictions based on the individual residues, and 0.79 for RONN. This indicates that local predicted RSA is strongly correlated with the disordered regions. This correlation is more significant than the correlation when no window is used. Our simple linear regression model, PredRSA9, shows comparable performance when compared with the RONN method. We observe that both methods perform similarly for low FP rates, while the regression model performs better for FP rates of above 0.2. Overall, we conclude that the predicted RSA constitutes a valuable input that could be used to predict both the B′-factor values and the disordered regions.

Our results show that B′-factor prediction model that was established using ordered residues can be used to predict disordered regions. At the same time, a disorder predictor like RONN that provides accurate predictions on the disordered regions may not yield high quality prediction of B-factor values, which implies that some other factor(s), besides B-factors, could be associated with the disorder.

Case studies

The discussed above relation is utilized in the context of case studies based on the observation that the protein structure, flexibility and function are closely linked.[6] Both rigid and flexible residues are important in implementing certain protein functions. For instance, enzyme active sites are in general rigid although most of them are located at the protein surface.[2] We consider three case studies that concern analysis of *Escherichia coli* ribonuclease HI (RNase HI),[38,78] and human interleukin-2 (hIL-2)[16] and CDK2[15] proteins. The first study concentrates on the analysis of rigid residues, while the other two studies concern flexible residues. In all the cases we investigate the relationship between residue flexibility, solvent accessibility, and protein function.

Rigid active site of RNase HI

RNase HI requires the binding of cofactors to perform its biochemical activity.[78] The active site residues, D10, E48, and D70 that bind Mg^{2+}, have been shown to be rigid (this was also shown by the sequence-based prediction method PROFbval[38]) and highly conserved.[78] At the same time, these residues are relatively solvent exposed with RSA values of 9.7%, 12.0%, and 40.2%, respectively, which are based on the calculation for the *apo* X-ray structure (PDB ID: 2RN2).

Figure 12 shows the actual B′-factor values along the RNase HI chain, as well as the predicted B′-factors that were obtained with our two linear regression models DsspRSA9 (that uses the knowledge of the structure) and PredRSA9 (that is based solely on the protein sequence), with WCN (which is a structure-based method),[24] and

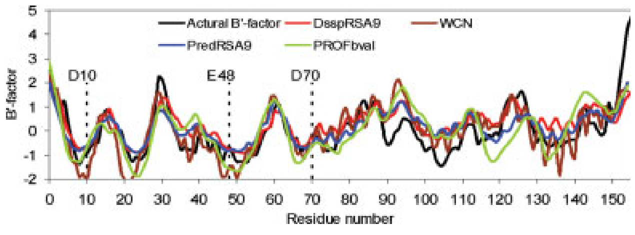

Figure 12

The actual B′-factors and the B′-factors predicted with DsspRSA9, PredRSA9, WCN and PROFbval for RNase HI (PDB ID: 2RN2). The residues that constitute the active site (D10, E48, and D70) are shown with dotted lines. The predictions of PROFbval were obtained using the web server.[38] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

with PROFbval (which is based on the sequence alone).[38] WCN values were computed based on the definition of WCN proposed in Ref. 24. We selected to use PROFbval since this method is readily available and has comparable quality to another sequence-based method by Yuan *et al.*,[29] which we could not obtain. The CC values between the actual the predicted B′-factor values equal 0.79, 0.75, 0.72, and 0.59 for DsspRSA9, WCN, PredRSA9, and PROFbval, respectively. This shows that our relatively simple linear model perform well when applied to predict B′-factor values.

All four predictors predicted this active site to be rigid. We observe that the adjacent (in the sequence) neighbors of the active site residues are relatively buried. More specifically, the RSA values of neighbors of D10 equal 1.4% for T9 and 5.1% for G11, the neighbors of E48 equal 0.0% for M47 and 0.0% for L49, and the neighbors of D70 equal 0.7% for T69 and 3.4% for S71. In the case of the most solvent-exposed residue in the active site, D70, we note that the burial of the adjacent residues coincides with the rigidity of this residue. This corroborates with our observation that two buried adjacent residues have promoting effect on the rigidity of the central-exposed residues.

We also compare the predictions made by using RSA of individual residues (DsspRSA1 and PredRSA1 models) with the predictions when using local RSA values (DsspRSA9 and PredRSA9 models). Application of the local RSA results in the increase of the correlation between the predicted and the actual B′-factor from 0.58 to 0.79 when using the actual RSA values (DsspRSA models) and from 0.54 to 0.72 when using the RSA predicted from the sequence (PredRSA models). This confirms the strong impact of neighboring residues on the relation between solvent accessibility and the flexibility of residues.

Flexible PEGylation site of human interleukin-2

Human interleukin-2 (hIL-2) is a pharmaceutical protein with a chimeric form that undergoes PEGylation mediated by tranglutaminase (TGase).[16] The flexibility or local unfolding of the chain region encompassing the Gln residue(s) was suggested as the main feature dictating the site-specific modification mediated by TGase.[16] It was found that the microbial TGase allows a selective and stoichiometric incorporation of the polyethyleneglycol (PEG) polymer chain at the Gln74 of hIL-2, in spite of the protein including six Gln residues in positions 11, 13, 22, 57, 74, and 126.[16] As shown in Figure 13, residue Gln74, which is adjacent to a disordered region (residue 75-76), is highly flexible based on its high B′-factor value (PDB ID: 1M47), while the other five Gln residues are relatively rigid. Fontana *et al.*[16] emphasize that the exposure to the surface is not sufficient to explain the site-specific TGase attack, since this protein has several surface-exposed Gln residues, while only one of them is located in a flexible region that is attacked by TGase.

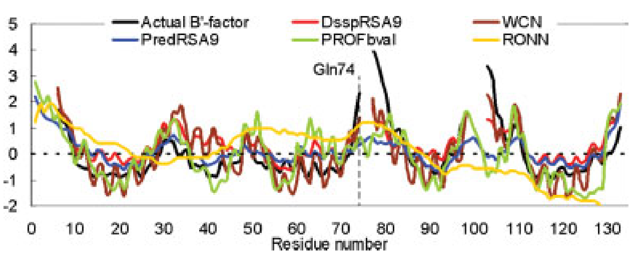

Figure 13

The actual B′-factors and the B′-factors predicted with DsspRSA9, PredRSA9, WCN, PROFbval and RONN for hIL-2 (PDB ID: 1M47). The site-specific PEGylation at Gln74 is shown with dotted line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 13 shows the B′-factor profiles of the X-ray structure of hIL-2 as well as the B′-factor values predicted by DsspRSA9, PredRSA9, WCN, PROFbval, and RONN methods. The CC values between the actual and the predicted B′-factor values equal 0.74, 0.72, 0.54, 0.48, and 0.19 for DsspRSA9, WCN, PredRSA9, PROFbval, and RONN, respectively. The former four methods predict Gln74 to be flexible (the predicted B′-factor is larger than zero) and encompassed in a flexible (disordered) region, while they predict the other five Gln residues to be rigid. The RONN method is used to find the disordered regions rather than to predict B′-factor values. The X-ray structure of 1M47 contains three disordered regions composed of residues 1–5, 75–76, and 99–102. While RONN performs well on the former two regions (the predicted values are relatively high), but it does not detect the third region. The sequence-based B′-factor predictors, PredRSA9, and PROFbval, succeed in identifying the three disordered regions (higher predicted B′-factor values correlate with location of the disordered regions), although the results are somehow weaker for the third region. At the same time, these two methods perform better in the context of the B′-factor prediction, while RONN is consistently characterized by lower CC with the actual B′-factor values.

Flexible regions of CDK2

CDK2 is the most thoroughly studied of the cyclin-dependent kinases that regulate essential cellular processes such as the cell cycle. The regulation and function of CDK2 have been intensively investigated.[15] The structural changes correlated with the intrinsic dynamics of this protein are shown to be essential for the successful execution of its biological functions.[3] Following Bártová *et al.*,[15] we investigated the functional flexibility of CDK2 based on four X-ray structures with PDB IDs 1HCK (free CDK2), 1FIN (CDK2/cyclin A/ATP), 1JST (pT160-CDK2/cyclin A/ATP), and 1QMZ (pT160-CDK2/

Table VI
The Correlation Coefficients Between the Actual and Predicted B′-Factor Values

Input	Methods	1HCK	1F1N	1JST	1QMZ
Structure	MD[a]	0.60	0.49	0.46	0.73
	GNM[a]	0.36	0.50	0.49	0.68
	WCN	0.55	0.59	0.51	0.61
	DsspRSA1	0.39	0.48	0.42	0.55
	DsspRSA9	**0.55**	**0.58**	**0.55**	**0.71**
Sequence	PROFbval[b]	0.38	0.42	0.40	0.50
	RONN	−0.10	0.12	−0.09	0.14
	PredRSA1	0.38	0.40	0.38	0.46
	PredRSA9	**0.47**	**0.44**	**0.47**	**0.56**

The values predicted with several methods, including MD, GNM, WCN, PROFbval, RONN, DsspRSA1, DsspRSA9, PredRSA1, and PredRSA9 for four CDK2 chains. The actual B-factor values were extracted from the chain A of 1HCK, 1F1N, 1JST, and 1QMZ. The results of the linear regression models that use local RSA values, which were developed in this paper, are shown in bold.

[a]The results were taken from Ref. 15.

[b]The predictions were obtained using the PROFbval web server.[38]

cyclinA/ATP/HHASSPRK). Table VI compares the CC values computed between the actual B′-factors and the B′-factors predicted with our linear regression models (DsspRSA1, PredRSA1, DsspRSA9, and PredRSA9) and other methods including structure-based methods such as MD,[15] GNM,[21] and WCN,[24] and sequence-based approaches such as PROFbval[30,38] and RONN.[39] Our structure-based method, DsspRSA9, provides comparable results with the results obtained with MD simulations, and better than results obtained with GNM and WCN. Using the sequence-based method results in the same prediction for all four chains (the chains are the same), although the correlation between the predicted and the actual B′-factors differ, we note that PredRSA9 shows the best performance when compared with PROFbval and RONN. Figure 14 shows the cartoon representations of the CDK2 (1HCK) colored by the actual B′-factors [Fig. 14(A)] and B′-factors predicted with DsspRSA9 [Fig. 14(B)]. We also note that the usage of local RSA improves the predictions when compared with using the RSA values for individual residues, that is, the results of DsspRSA9 and PredRSA9 are better than the results of DsspRSA1 and PredRSA1, respectively.

Table VII lists flexible regions, which are associated with peaks in the B′-factor profile and which were discussed by Bártová *et al.*,[15] and describes their functional roles in CDK2. These peaks are also visualized in Figure 14(B). We added the functional descriptions for regions *d* and *f*, since the authors in Ref. 15 labeled them without explaining their functional roles. We examine whether the B′-factor profiles predicted by the proposed linear regression models and other structure- and sequence-based methods also include peaks at the positions corresponding to the functional sites. The actual B′-factor profiles and the profiles computed with MD, GNM, WCN, PROFbval, RONN, DsspRSA9, and PredRSA9 are given in Figure 15. For consistency, the results of MD, GNM, which were received from the authors of Ref. 15, WCN, and RONN were normalized in the same way as the original B-factors were normalized. Table VII shows whether the profile generated each of the considered methods includes a peak at the positions corresponding to the functional sites. We assume that the peak corresponds to a local maximum for which the B′-factor value is >0.

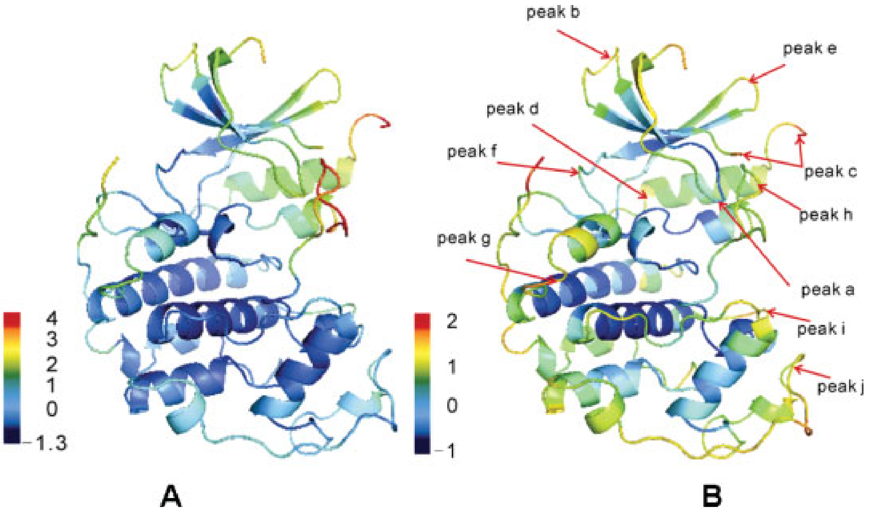

Figure 14
The cartoon representations of the CDK2 (1HCK) colored by (**A**) actual B-factors and (**B**) B-factors predicted using DsspRSA9. The color scale ranges between red which denotes highly flexible regions, through yellow and green that depicts moderately flexible regions, and light and dark blue that correspond to rigid regions. Panel B also annotates location of flexible regions that are denoted as peak *a* through peak *j*.

Table VII
Flexible Regions in CDK2 Family[15]

		Method						
Peak	Function description	DsspRSA9	WCN	MD	GNM	PredRSA9	PROFbval	RONN
a: 13–15	Glycine-rich loop (G-loop) as an inhibitory segment of CKD2 roofing the ATP-binding site.	+	+	+				
b: 25	A hinge between the secondary structure elements	+	+	+	+	+	+	
c: 36–42	Provides the flexibility of the α1-helix required to allow its shift during activation, i.e., after binding of the regulatory subunit.	+	+	+	+	+	+	+
d: 57	A conserved arginine R122 that is highly buried upon binding to cyclin forms a salt bridge with the highly conserved glutamate E57 within the CDK2 family.[79]	+	+	+	+	+	+	
e: 71–76	This region contacts the regulatory subunit and presumably also CDK2 substrates	+	+	+	+	+	+	
f: 84–85	H84 carbonyl and Q85 CA shift toward the inhibitor in the active CDK2.[80]							
g: 95–98	The loop between the α2 and α3 helices; its flexibility increases when CDK2 is in complex with cyclin A, while α3-helix flexibility decreases after cyclin A binding.	+	+	+	+	+	+	
h: 150–162	The activation segment (T-loop) which is flexible in CDK2 that is not bound to cyclin but not flexible in other CDK2–cyclin complexes; in cases where cyclin A is present, this peak vanishes because of the intensive interaction between the T-loop and cyclin A.	+	+	+	+	+	+	
i: 177–180	The highly flexible region in all of the inactive, semi-active, and active forms of CDK2.	+	+	+		+	+	
j: 220–260	This region includes nearly all of the highly mobile CMGC insert (also called the ``CDK insert`` region that comprises of residues 219–251).	+	+	+	+	+	+	+

The first column shows flexible regions that correspond to peaks in the B′-factor profiles; the second column provides a brief description of the regions; the third column shows whether the B′-factor profile predicted with a given method has a peak in the corresponding region (+ denotes that the peak was found, while empty cell denotes lack of the peak).

We observe that the results generated by DsspRSA9 are comparable to results obtained with the other three structure-based methods, MD, GNM, and WCN. The four methods generated strong peaks for the regions *b, c, e, g, h*, and *j*. Similarly as MD, WCN, and DsspRSA9 identified all peaks except for the region *f*, whereas GNM did not generate the peaks for the regions *a* and *f*. Visual inspection of the original B′-factor profile reveals that some peaks, such as for the regions *d, f*, and *i*, are relatively weak, and thus they are more difficult to predict. In the case of the peak at the region *i*, MD, WCN, and DsspRSA9 reflected it relatively better than GNM, while the four methods perform similarly for the regions *d* and *f*. When considering the B′-factor value in the region *b*, DsspRSA9 provides predictions that are the closest to the actual values when considering all four CDK2 chains. More specifically, only the one HCK chain, that is, the free CDK2, has a peak in this position, which is also true for the predictions of DsspRSA9, while MD, WCN, and GNM predict the peaks for all the four chains.

The sequence-based methods, that is, PredRSA9, PROFbval, and RONN, yield one prediction for CDK2 because all the four chains have the same sequence. At the same time, their performance with respect to the detection of the peaks is quite different. The two methods that are designed to compute B′-factor values, PredRSA9 and PROFbval, captured almost all flexible regions except for the regions *a* and *f*. The RONN method that aims at the prediction of disorder regions found only two peaks in the *c* and *j* regions. The region *c*, which concerns residues 36 to 42, is disordered in several X-ray structures in the noncomplexed CDK2 proteins, e.g., 1AQ1, 1CKP, 1HCK, 1HCL. This region has "dual personality," it becomes ordered in the active CDK2 in the complex.[81] This is likely the reason that RONN perform well in the case of this peak.

The above three case studies demonstrate that our relatively simple models that use solvent accessibility to compute the B′-factor values can be used to infer positions of functionally important residues, and the quality of such information is at least comparable with the quality of information provided by other modern methodologies.

DISCUSSION

Our analysis suggests that the local solvent accessibility strongly influences the residue flexibility. The correlations

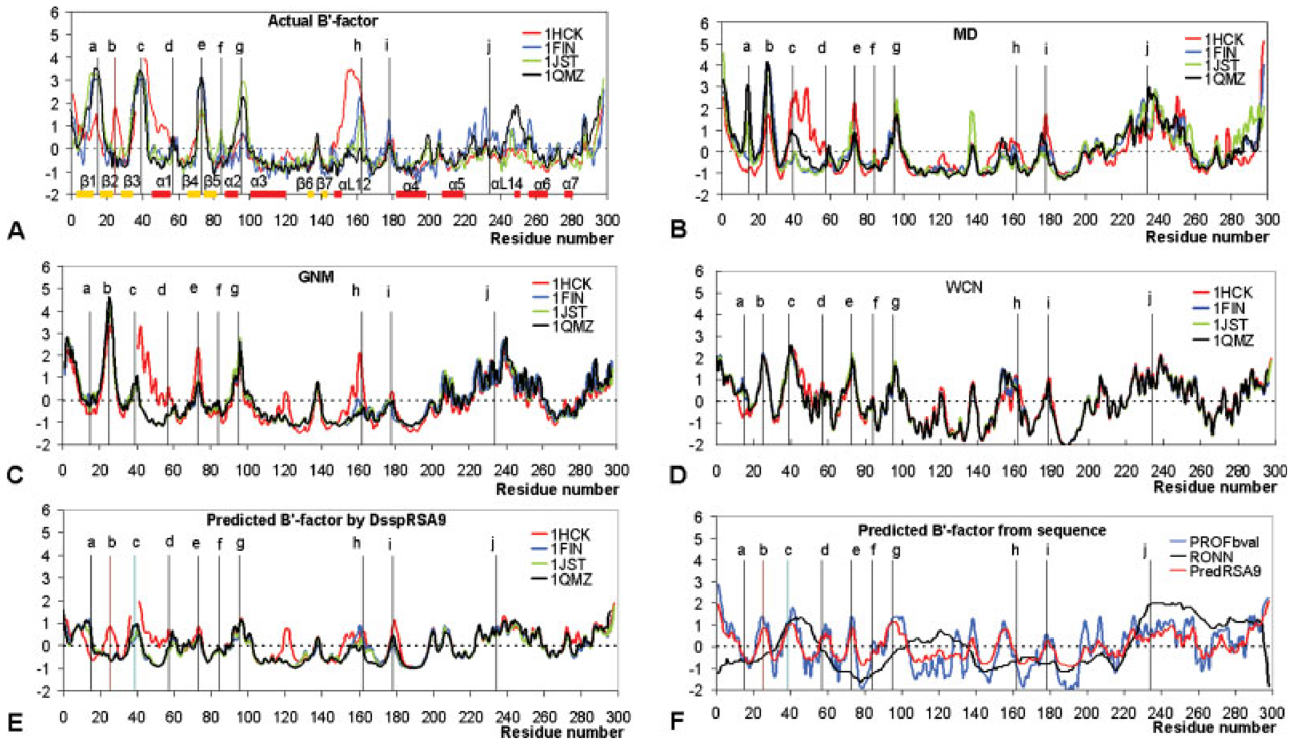

Figure 15

Comparisons of B′-factor profiles from (**A**) the X-ray crystal structures, (**B**) MD, (**C**) GNM, (**D**) WCN, and (**E**) DsspRSA9. The profile for 1HCK (free CDK2) is shown in red, for 1FIN (CDK2/cyclin A/ATP) in blue, for 1JST (pT160-CDK2/cyclin A/ATP) in green, and for 1QMZ (pT160-CDK2/cyclin A/ATP/HHASSPRK) in black. Panel (**F**) compares B′-factor values of CDK2 predicted by the sequenced-based methods including PredRSA9, PROFbval, and RONN. Regions discussed in Table VII are labeled by the vertical lines where (a) corresponds to residue 15, (b) to residue 25, (c) to residue 38, (d) to residue 57, (e) to residue 73, (f) to residue 84, (g) to residue 95, (h) to residue 162, (i) to residue 178, and (j) to residue 234. The B′-factor profiles are missing for region *c* in the case of panels A, C, D, and E since this region is disordered, which implies lack of the atomic coordinates. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

that were obtained based on our regression models should be considered as relatively strong in the context of the noise imbedded in the experimental B-factor values. The B-factors not only reflect the fluctuation and static, dynamic and lattice disorders, but also depend on the experimental resolution, crystal contacts and refinement procedures. Examination of the B-factor profiles of homologous proteins shows that they are correlated with each other with the average CC of 0.80.[28,29] This constitutes an approximate upper limit when it comes to prediction of B-factor values, which also applies to our regression model. The high ACC of B-factor predictions that were obtained from RSA, that is, 0.66 when using the actual RSA values and 0.55 with the RSA values predicted from the protein sequence, suggests that, to some extent, the RSA-based models could be used to derive the information concerning residue flexibility. This information, in turn, has implication in characterization of protein structure–function relation, which we demonstrate in our three case studies.

To date, several *in silico* methods, which use either protein structure or sequence as the input, were developed to describe the flexibility and/or to predict the flexibility at the residue, region, or protein level. GNM[21] regards a protein as an elastic network of C_α atoms of which the fluctuations are assumed to be isotropic and follow Gaussian distributions. WCN[24] is established based on the inter-residue contacts weighted by the distances among residues. These two structure-based methods show high correlations with the experimental B-factors and are useful to characterize the cross-correlations of motions among residues. We note that the proposed RSA based models, DsspRSA and PredRSA, cannot be used to perform the cross-correlation analysis. At the same time, in spite of its simplicity our structure-based DsspRSA model shows higher correlation with the experimental B-factors when compared with the above two structure-based methods. The advantage of the sequence-based PredRSA model is that it can be used for high throughout analysis of proteins for which structural coordinates of residues are unavailable.

CONCLUSIONS

We investigate the relationship between the residue flexibility, measured by B-factor, and the solvent accessibility in the context of the influence of information in local sequence neighborhood, different residue types, and different secondary structures. The main findings are:

- The ACC between B-factor and the actual RSA equals 0.52 which indicates relatively strong relation. The FECI that measures CC between B-factor and RSA for the 20 amino acids is highly correlated with the stability scale derived from the knowledge-based atom–atom potentials that characterizes the average contributions of each amino acid to the folding stability.
- The exposure or burial of the adjacent residues strongly influences (promotes or inhibits) the flexibility of the central residues. Our results, which are independent of the RSA thresholds used to define buried/exposed residues, suggest that the exposed residues with two buried adjacent residues have lower mean B′-factor (are more rigid) than the buried residues with two exposed neighbors.
- Using a linear regression model we observed that the inclusion of local RSAs significantly improves the correlation between the solvent accessibility and the B-factor. This increase is consistent for all amino acid type and the underlying secondary structures. The strength of the relation decreases linearly, with a factor of 0.5, with the distance from the considered residue.

We also contrasted the relation between the flexibility and RSA with the relation between flexibility and distance/volume based residue depth. When considering individual residues, the volume-based depth has the strongest correlation with flexibility. Inclusion of the local information (local RSA/depth) results in significant increase of the correlation with RSA, which becomes stronger than the correlation with the depth.

Furthermore, we observe that the RSA that is predicted from the sequence could be used to distinguish between the disordered and ordered residues by utilizing information about the B′-factor that is computed (predicted) from the predicted RSA values. Similarly as in the case of the actual RSA values, inclusion of the local predicted RSA values helps in providing a better contrast between the disordered and the ordered residues. Our results suggest that the B-factor values predicted using the RSA values could be used to identify the flexible and rigid regions and also to find the disordered regions.

Prediction models developed based on the local actual RSA (structure-based) and the local predicted RSA (sequence-based) show similar or better results in the context of B-factor and disordered/ordered residues predictions when compared with several competing approaches on three large benchmark datasets and three case studies. The proposed linear regression-based models provide an interesting insight into the structure–flexibility relation (DsspRSA model) and the sequence–flexibility relation (PredRSA model), which in both cases can be extended into applications in the context of the structure/sequence–flexibility–function relation.

ACKNOWLEDGMENTS

The authors thank Dr. Yuan for clarification concerning the normalization of B-factor values, Drs. Dor and Zhou for explaining their Real-SPINE program, and Dr. Esnouf for providing the RONN program. The authors are also grateful to Drs. Bártová and Otyepka for providing the MD and GNM results for CDK2.

REFERENCES

1. Karplus M, McCammon JA. The internal dynamics of globular proteins. CRC Crit Rev Biochem 1981;9:293–349.
2. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109–114.
3. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci USA 2005;102:18908–18913.
4. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D. Intrinsic dynamics of an enzyme underlies catalysis. Nature 2005;438:117–121.
5. Dodson G, Verma CS. Protein flexibility: its role in structure and mechanism revealed by molecular simulations. Cell Mol Life Sci 2006;63:207–219.
6. Bhalla J, Storchan GB, MacCarthy CM, Uversky VN, Tcherkasskaya O. Local flexibility in molecular function paradigm. Mol Cell Proteomics 2006;5:1212–1223.
7. Vihinen M. Relationship of protein flexibility to thermostability. Prot Eng 1987;1:477–480.
8. Parthasarathy S, Murthy MRN. Analysis of temperature factor distribution in high-resolution protein structures. Protein Sci 1997; 6:2561–2567.
9. Carugo O, Argos P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. Proteins 1998;31:201–213.
10. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–734.
11. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Bio 2004;338:181–199.
12. Kurgan L, Cios KJ, Zhang H, Zhang T, Chen K, Shen S, Ruan J. Sequence-based methods for real value predictions of protein structure. Curr Bioinformatics 2008;3:183–196.
13. Yang CY, Wang R, Wang S. M-score: a knowledge-based potential scoring function accounting for protein atom mobility. J Med Chem 2006;49:5903–5911.
14. Sandhu KS, Dash D. Conformational flexibility may explain multiple cellular roles of PEST motifs. Proteins 2006;63:727–732.
15. Bártová I, Koca J, Otyepka M. Functional flexibility of human cyclin-dependent kinase-2 and its evolutionary conservation. Protein Sci 2008;17:22–33.
16. Fontana A, Spolaore B, Mero A, Veronese FM. Site-specific modification and PEGylation of pharmaceutical proteins mediated by transglutaminase. Adv Drug Deliv Rev 2008;60:13–28.
17. Pandey BP, Zhang C, Yuan X, Zi J, Zhou Y. Protein flexibility prediction by an all-atom mean-field statistical theory. Protein Sci 2005;14:1772–1777.

18. Rueda M, Ferrer-Costa C, Meyer T, Pérez A, Camps J, Hospital A, Gelpí JL, Orozco M. A consensus view of protein dynamics. Proc Natl Acad Sci USA 2007;104:796–801.
19. Yang L, Song G, Jernigan RL. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys J 2007;93:920–929.
20. 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.
21. 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.
22. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
23. Lu CH, Huang SW, Lai YL, Lin CP, Shih CH, Huang CC, Hsu WL, Hwang JK. On the relationship between the protein structure and protein dynamics. Proteins 2008;72:625–634.
24. Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. Deriving protein dynamical properties from weighted protein contact number. Proteins 2008;72:929–935.
25. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72:212–213.
26. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
27. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12:1060–1072.
28. 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.
29. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein b-factor profiles. Proteins 2005;58:905–912.
30. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005;61:115–126.
31. Yang LW, Bahar I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 2005;13:893–904.
32. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
33. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J 2007;92:1439–1456.
34. Lobley A, Swindells MB, Orengo CA, Jones DT. Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 2007;3:e162.
35. Ferron F, Longhi S, Canard B, Karlin D. A Practical overview of protein disorder prediction methods. Proteins 2006;65:1–14.
36. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69 (Suppl 8):129–136.
37. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6. Proteins 2005;61 (Suppl 7):167–175.
38. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 2006;22:891–893.
39. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21:3369–3376.
40. Worch R, Stolarski R. Stacking efficiency and flexibility analysis of aromatic amino acids in cap-binding proteins. Proteins 2008;71:2026–2037.
41. Dunker AK, Obradovic Z. The protein trinity-Linking function and disorder. Nat Biotechnol 2001;19:805–806.
42. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
43. Connoly ML. Solvent accessibility surfaces of protein and nucleic acids. Science 1983;221:709–713.
44. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
45. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information in the prediction of protein stability changes, comparison between buried and partially buried mutations. Protein Eng 1999;12:549–555.
46. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
47. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 2007;68:636–664.
48. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
49. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
50. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;57:558–564.
51. Wang JY, Lee HM, Ahmad S. SVM-cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. Proteins 2007;68:82–91.
52. Sheriff S, Hendrickson WA, Stenkamp RE, Sieker LC, Jensen LH. Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. Proc Natl Acad Sci USA 1985;82:1104–1107.
53. Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Eng 1997;10:777–787.
54. Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2:527–541.
55. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 1999;7:723–732.
56. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J 2003;84:2553–2561.
57. Varrazzo D, Bernini A, Spiga O, Ciutti A, Chiellini S, Venditti V, Bracci L, Niccolai N. Three-dimensional computation of atom depth in complex molecular structures. Bioinformatics 2005;21:2856–2860.
58. Pedersen TG, Sigurskjold BW, Andersen KV, Kjaer M, Poulsen FM, Dobson CM, Redfield C. A nuclear-magnetic-resonance study of the hydrogen-exchange behavior of lysozyme in crystals and solution. J Mol Biol 1991;218:413–426.
59. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J 2004;86:85–91.
60. 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.
61. 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.
62. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.
63. Altschul SF, Madden TL, Schäffer AA, Zhang J, 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.
64. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
65. Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 2007;68:76–81.
66. Yuan Z, Wang ZX. Quantifying the relationship of protein burying depth and sequence. Proteins 2008;70:509–516.
67. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan LA. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinformatics 2008;9:388.

68. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 1996;38:305–320.
69. Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann; 2005. 525 p.
70. Team RDC. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2007.
71. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374.
72. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004;54:315–322.
73. Klein-Seetharaman J, Oikawa M, Grimshaw SB, Wirmer J, Duchardt E, Ueda T, Imoto T, Smith LJ, Dobson CM, Schwalbe H. Long-range interactions within a nonnative protein. Science 2002;295: 1719–1722.
74. Zhou R, Eleftheriou M, Royyuru AK, Berne BJ. Destruction of long-range interactions by a single mutation in lysozyme. Proc Natl Acad Sci USA 2007;104:5824–5829.
75. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;(Suppl 5):192–199.
76. Sonego P, Kocsor A, Pongor S. ROC analysis: applications to the classification of biological sequences and 3D structures. Brief Bioinform 2008;9:198–209.
77. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940–3941.
78. Goedken ER, Keck JL, Berger JM, Marqusee S. Divalent metal cofactor binding in the kinetic folding trajectory of *Escherichia coli* ribonuclease HI. Protein Sci 2000;9:1914–1921.
79. Kannan N, Neuwald AF. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK. CDK, GSK, SRPK, DYRK, and CK2α Protein Sci 2004;13:2059–2077.
80. Kontopidis G, McInnes C, Pandalaneni SR, McNae I, Gibson D, Mezna M, Thomas M, Wood G, Wang S, Walkinshaw MD, Fischer PM. Differential binding of inhibitors to active and inactive CDK2 provides insights for drug design. Chem Biol 2006;13:201–211.
81. Zhang Y, Stec B, Godzik A. Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. Structure 2007;15:1141–1147.