BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 19 2009, pages 2552–2558
doi:10.1093/bioinformatics/btp463

Structural bioinformatics

Methyl side-chain dynamics prediction based on protein structure

Pablo Carbonell[1,2] and Antonio del Sol[1,*]

[1]Fujirebio Inc., Bioinformatics Research Unit, Research and Development Division,Komiya-cho, Hachioji-shi, Tokyo 192-0031, Japan and [2]University of Evry, Epigenomics Program, Genopole, 523 Terrasses de l'Agora, 91034 Évry cedex, France

Received on March 5, 2009; revised on July 4, 2009; accepted on July 23, 2009

Advance Access publication July 30, 2009

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein dynamics is believed to influence protein function through a variety of mechanisms, some of which are not fully understood. Thus, prediction of protein flexibility from sequence or structural characteristics would assist in comprehension of the ways dynamics is linked to function, and would be important in protein modeling and design. In particular, quantitative description of side-chain dynamics would allow us to understand the role of side-chain flexibility in different functional processes, such as protein–ligand and protein–protein interactions.

Results: Using a dataset of 18 proteins, we trained a neural network for the prediction of methyl-bearing side-chain dynamics as described by the methyl side-chain generalized order parameters (S^2) inferred from NMR data. The network uses 10 input parameters extracted from 3D structures. The average correlation coefficient between the experimental and predicted generalized order parameters is $r = 0.71 \pm 0.029$. Further analysis revealed that the order parameter depends more strongly on the methyl carbon packing density, the methyl carbon distance to the C_α atom, and the knowledge-based pair-wise contact potential between the methyl carbon and neighboring amino acids. In general, we observed an improvement in the prediction of methyl order parameters by our network in comparison with molecular dynamics simulations. The sensitivity of the predictions to minor structural changes was illustrated in two examples (calmodulin and barnase) by comparing the S^2 predictions for the unbound and ligand-bound structures. The method was able to correctly predict most of the significant changes in side-chain dynamics upon ligand binding, and identified some residues involved in long-range communications or protein–ligand binding.

Availability: http://epigenomique.genopole.fr/∼carbonell

Contact: antdelsol@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Proteins molecules are not rigid entities, but undergo different types of motion covering a wide range of time scales and amplitudes. It is now well accepted that protein dynamics plays a key role in protein function, including enzyme catalysis, ligand recognition, protein–protein interactions and allosteric communications (Goodey and Benkovic, 2008; Henzler-Wildman and Kern, 2007; Igumenova *et al*., 2006). Although an increasing number of experimental and theoretical studies have shed some light on the manner in which protein motions are related to molecular function, this relationship is not fully understood and is a subject of intense controversy (Goodey and Benkovic, 2008; Henzler-Wildman and Kern, 2007; Igumenova *et al*., 2006). Thus, the elucidation of structural and sequence characteristics determining protein conformational flexibility would assist the understanding of protein function and would be relevant for protein modeling and design.

Here, we focus on protein side-chain dynamics, which is more complex and heterogeneous than that of the backbone, and has been shown to be a major component of protein conformational entropy (Frederick *et al*., 2007; Trbovic *et al*., 2009; Yang and Kay, 1996). Indeed, experimental results show that changes in side-chain dynamics are associated with protein–protein or protein–ligand interactions in cases with minimal structural changes, indicating the potential role of side-chain motions in modulating binding affinity and propagating long-range signals within a protein (Boyer and Lee, 2008; Frederick *et al*., 2007; Fuentes *et al*., 2004; Namanja *et al*., 2007). NMR relaxation measurements of the side-chain methyl atoms have been routinely used to probe side-chain mobility (Boyer and Lee, 2008; Frederick *et al*., 2007; Fuentes *et al*., 2004; Namanja *et al*., 2007). Indeed, there is a great interest in obtaining information about methyl side-chain dynamics, since methyl groups are frequently found at many different sites throughout a protein, including protein–protein/ligand binding interfaces, as well as inside protein hydrophobic cores (Jones *et al*., 1976; Nicholson *et al*., 1992). Therefore, information on dynamics of methyl groups provides a good probe for side-chain motion at different protein sites. For example, Lee and co-workers making the assumption that non-methyl bearing amino acids respond on average as methyl bearing residues in calmodulin (Cam) upon smMLCKp domain binding, estimated the change in conformational entropy of Cam based on the change in dynamics (Lee *et al*., 2000), and their result agrees remarkably well with the estimation by Wintrode and Privalov (1997) from calorimetric measurements. The relaxation data, which explore side-chain motions in the picosecond-to-nanosecond time regime, can be further interpreted in the form of Lipari-Szabo S^2 order parameters and τ_e internal correlation times

[*]To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

(Lipari and Szabo, 1982). In this case, the S^2 order parameter, which takes values between 0 and 1, is a measure of the degree of spatial restriction of a given methyl group.

Different authors have proposed theoretical models for representing side-chain flexibility. Molecular dynamics simulations have been used for estimating side-chain order parameters (Best *et al*., 2005; Hu *et al*., 2005; Prabhu *et al*., 2003; Showalter *et al*., 2007). A Monte-Carlo-based approach has recently been proposed to model side-chain variability in protein design simulations (Friedland *et al*., 2008). Simple models have also been introduced to predict side-chain order parameters based on the methyl carbon packing density or solvent accessibility (Mittermaier *et al*., 1999, 2003). A different approach for modeling side-chain dynamics considered the structural variability obtained from the analysis of different conformers of the same protein (Best *et al*., 2006). However, although some of these methods achieve considerable accuracy, they are generally time-consuming and often computationally intractable for the analysis of a large number of proteins. On the other hand, methyl side-chain order parameters are difficult to predict based on individual sequence or structural characteristics. Indeed, it has been shown that individual parameters do not correlate or poorly correlate with side-chain methyl order parameters (Igumenova *et al*., 2006), suggesting that side-chain mobility depends on several factors in a complex manner.

In the present work, we trained a specific type of artificial neural network (multi-layered feedforward back-propagation network) for the prediction of methyl side-chain generalized order parameters (S^2) based only on the 3D structure; an approach which has been similarly used for the prediction of backbone dynamics (Trott *et al*., 2008). We compiled a dataset of 2697 experimentally determined methyl side-chain order parameters (18 proteins) from the Biological Magnetic Resonance Bank and literature (Seavey *et al*., 1991); and retrieved the structures of the corresponding proteins from the Protein Data Bank (PDB) (Berman *et al*., 2000). The average Pearson's correlation coefficient between experimental and predicted values of the generalized order parameters is $r = 0.71 \pm 0.029$. Further investigations revealed that the order parameter depends more strongly on the methyl carbon packing density, the methyl carbon distance to the C_α atom and the knowledge-based pair-wise contact potential between the methyl carbon and neighboring amino acids. Some of these characteristics were already considered in a simple analytical model proposed by Ming and Brüschweiler (2004); however, a significant improvement in prediction is obtained using our neural network in comparison with their model. This finding indicates that side-chain dynamics is conditioned by a set of different characteristics, rather than single features. Furthermore, the predictions of methyl order parameters by molecular dynamics simulations for five proteins (ubiquitin, TNfn3 domain, FNfn10 domain, barnase and Cam) were, in general, improved by our method. In order to illustrate the sensitivity of our predictions to details of the protein structure, we selected two examples of proteins (Cam and barnase) exhibiting changes in side-chain dynamics and minor structural changes upon ligand binding. Our predictions of residues with significant changes in methyl side-chain generalized order parameters upon ligand binding are in good agreement with the experimental data, and several of these amino acids have been previously reported as residues involved in long-range communications or participating in the protein–ligand interface. Thus, our neural network can be used as a predictive tool for detecting significant perturbations of side-chain dynamics upon a change in functional state, and can complement other theoretical methods in drug design.

2 METHODS

2.1 The dataset

We compiled a dataset of methyl side-chain order parameter values (S^2-values) from the Biological Magnetic Resonance Bank (BMRB) (Eldon *et al*., 2007), and from some other published NMR experiments, totaling 2697 entries obtained in 56 experiments from 18 non-redundant proteins (see Supplementary Table S1 for details). Table 1 shows the distribution of the dataset according to amino acid type and methyl group. In order to improve the data quality for training the neural network, the original dataset was filtered to remove those S^2-values with experimental error >0.05. The filtered dataset contained 2488 entries.

There were nine examples in the dataset of proteins in two or more different states: Cam, barnase, pdz, mup, folate, staphylococcal nuclease, pin1, sh3 domain and MSG.

2.2 Feature set

Side-chain methyl groups are represented by 10 features, which were extracted from the protein 3D structure deposited in the PDB database. These features can be classified in two groups: geometry-based parameters; and knowledge-based potentials. Characteristics were measured at residue level in six out of these 10 parameters; whereas the rest (packing, elongation, pair-wise contact potential and rotameric state) are specific to each methyl group in the side chain.

2.2.1 Geometry-based parameters There are four parameters which evaluate motional restrictions of the methyl group in the side chain; while three additional parameters measure restrictions at the backbone.

(1) Packing density of methyl carbons: here, we introduced a modification to the contact sum C_i defined by Ming and Brüschweiler (2004). Namely, each contributing contact between the methyl carbon and a neighbor heavy atom is weighted by the packing of the later. The contact sum of methyl carbon *i* was given by:

$$C_i = \sum_j e^{-r_{ij}}, \tag{1}$$

where r_{ij} is the distance between the methyl carbon atom *i* and heavy atom *j*, for any contact other than with the methyl carbon side chain atoms, and within the limits of a sphere of radius 5 Å. Based on this

Table 1. Number of entries in the dataset

Amino acid	Methyl group	S^2-values	S^2-values (error ≤ 0.05)
ALA	CB	324	260
ILE	CD1	320	309
ILE	CG2	253	233
LEU	CD1	380	342
LEU	CD2	351	320
MET	CE	179	177
THR	CG2	248	225
VAL	CG1	322	310
VAL	CG2	320	312
Total		2697	2488

S^2-values for each type of amino acid and methyl group in the dataset.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

expression, we define the packing density of methyl carbon i as:

$$P_i = \sum_j C_j e^{-r_{ij}}. \qquad (2)$$

(2) Side chain stiffness: this parameter is the sum of packing densities as in Equation (2) for consecutive heavy atoms in the side chain connecting the methyl carbon i with the backbone, weighted by the relative number of dihedral angles separating the heavy atom from the backbone:

$$S_i = \sum_{j=1}^{N_i} jP_j, \qquad (3)$$

where j goes through the number of consecutive dihedral angles from the backbone and N_i is the total number of consecutive dihedral angles separating the backbone from the methyl carbon i.

(3) Elongation: this measure is defined as the distance of the methyl group from the C_α

$$E_i = r_{i\alpha}. \qquad (4)$$

(4) Side chain rotameric distance: we performed a statistical analysis of rotameric states through all methyl groups in PDB structures. The result was summarized in a table of dihedral angles χ representing the most common rotameric states (Supplementary Table S7) (Lovell *et al*., 2000). For each side chain, we measured the angular distance to its closest common rotameric state:

$$\Delta\chi = \chi - \chi_0. \qquad (5)$$

where χ_0 is the closest common rotameric state.

(5) Carbonyl backbone packing density: the definition is analogue to Equation (2), computed in this case for the carbonyl group at the backbone.

(6) Amide backbone packing density: the definition is analogue to Equation (2), computed in this case for the amide group at the backbone.

(7) Backbone hydrogen bonds: this parameter counts the number of hydrogen bonds involving backbone atoms, computed from the DSSP database (Kabsch and Sander, 1983).

2.2.2 *Knowledge-based potentials*

(1) Pairwise contact potential: we derived a pair-wise contact potential between the methyl carbon i and each amino acid type defined as:

$$K_i = -\sum_j \log\left(f_j\left(r_{ij}\right)\right). \qquad (6)$$

where r_{ij} is the distance between the carbon in the methyl group i and the closest heavy atom in the contacting j amino acid, and $f_j(r_{ij})$ is the frequency of occurrence of such contact for this type of amino acid and distance r_{ij} in the PDB database, with a bin interval of 0.5 Å (Supplementary Table S8).

(2) Van der Waals effects: this parameter provided a computation of Van der Waals interactions of methyl groups with its surrounding atoms.

(3) Solvation effects: this parameter was based on DSSP computation of solvent accessibility for each residue weighted by its hydrophobicity.

2.3 Artificial neural network

We used a multi-layered feedforward back-propagation network with one hidden layer (Haykin, 1998) for the prediction of methyl side-chain order parameters. Each unit in the input layer was fed with one of the 10 parameters in the feature set; hidden units were built with symmetrical sigmoids; and output layers with standard sigmoids, which gave a prediction of the S^2-value for the methyl group. The number of hidden units was empirically determined to be twice the number of input parameters. The network was implemented using the Fast Artificial Neural Network Library (FANN) (Nissen, 2003).

A neural network ensemble was separately trained for each methyl group belonging to different amino acid types. Special consideration was taken in the case of amino acids containing two methyl groups. Due to its symmetry, one network was used for both methyls in leucine and valine. In the case of isoleucine, the inputs of the neural network consisted of a combination of the two side-chain methyl group parameters, and the output was formed by the prediction of both S^2-values.

The neural network was trained by using the back propagation algorithm, with a stopping threshold for the sum of squared errors of $\varepsilon \leq 0.005$, which was empirically determined to provide a good tradeoff between model uncertainty and network overfitting.

2.4 Model validation

The validation of the neural network predictions was performed by means of a 500-fold sub-sampling cross-validation, where 80% of the dataset was used for training and 20% for validation. After the training of the neural network, the correlation coefficient between the prediction and the experimental data for the validation subset was sampled, being finally the test performed over the distribution of correlation coefficients. The validation was averaged over 10 trained networks for each random generation, in order to make the validation more independent of the convergence of the training set.

2.5 Protein-based cross-validation

The training set for each example was formed by the protein structures contained in our dataset, except for the protein under study and its homologues representing different conformational states. The validation set consisted of those methyl groups with accurate experimental S^2-values for the protein in the analyzed state. Validation was performed only for those proteins with experimental values available for at least 50% of the methyl groups (Supplementary Table S1).

3 RESULTS

3.1 Distribution of methyl side-chain S^2-values

Methyl side-chain S^2-values in the dataset follow a trimodal distribution (Supplementary Table S12, Fig. S1), which is consistent with previous observations (Lee and Wand, 2001). Furthemore, backbone dynamics was observed to be in general lower than side-chain dynamics (Lee *et al*., 2000), with a mean value of $S^2 = 0.85 \pm 0.14$; whereas for methyl side-chains the mean value was found to be $S^2 = 0.60 \pm 0.23$. Both values are almost uncorrelated with a correlation coefficient of $r = 0.25$ (Supplementary Fig. S2). Moreover, dynamics of the two methyl groups in leucine and valine are highly correlated (Mittermaier *et al*., 1999), with correlation coefficients of $r = 0.88$ and 0.94 respectively, whereas no strong correlation ($r = 0.63$) was found between the two methyl groups in isoleucine, which are asymmetrically located (Supplementary Fig. S3).

3.2 Influence of features on side-chain dynamics

It has been recognized by several authors that there are no simple structural determinants of methyl dynamics in proteins (Igumenova *et al*., 2006). In order to test the influence of our parameters on side-chain methyl dynamics, we checked the correlation of each single feature with the experimentally measured order parameters. Table 2 shows the correlation coefficients between experimental S^2-values and input features. Most of the features are weakly correlated with the order parameter. The highest correlations were found for elongation (distance of the methyl group to the C_α), pair-wise contact potential, and methyl packing density.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

Table 2. Correlation coefficients and P-values between experimental S^2-values and input features

Packing	Stiffness[a]	Elongation	Carbonyl	Amide
0.19 1.2×10^{-21}	-0.13[a] 7.6×10^{-11}	-0.30 6.5×10^{-53}	0.16 9.9×10^{-16}	-0.03 1.3×10^{-1}
hbonds	$\Delta\chi$[b]	Potential	vdw	Solvation
0.09 6.9×10^{-6}	0.05 1.2×10^{-2}	-0.20 7.3×10^{-24}	-0.00 9.9×10^{-1}	-0.14 2.3×10^{-12}

[a] Stiffness was computed excluding alanine.

[b] Averaged over corresponding methyl groups in Supplementary Table S7.

Correlation coefficients between S^2-values and input features for each amino acid type can be found in the Supplementary Tables S2, S3 and S4, as well as a detailed list of the most influential parameters for each amino acid is in Supplementary Table S5.

Furthermore, we computed the covariances between all parameters (Supplementary Table S6). The highest covariances were found in the following cases: between side-chain elongation and stiffness (cov ~ 0.68); between solvent effects and pair-wise potential (cov ~ 0.45); and between solvent effects and backbone hydrogen bonds (cov ~ −0.41).

Multiple regression fitting for linear and second-order polynomial models were not significant, with $R^2 = 0.16$ and 0.34, respectively, suggesting that our neural network reproduces a more complex non-linear dependence between the input parameters and the output.

3.3 Validation of the neural network

3.3.1 Random cross-validation We performed a 500-fold sub-sampling cross-validation of the neural network on the dataset of 18 proteins with experimental information on methyl side-chain order parameters (Section 2). The correlation coefficients in our test follow a normal distribution (Supplementary Fig. S4), with an average correlation coefficient $r = 0.71 \pm 0.029$ (P-value $= 4.6 \times 10^{-87}$). For the same dataset, using the Ming–Brüchsweiler's prediction method (Ming and Brüchsweiler, 2006) we obtained a correlation coefficient $r = 0.56$.

3.3.2 Protein-based cross-validation Each protein in the dataset with a significant number of accurate experimental values was cross-validated against the rest of the dataset. Correlation coefficients obtained for each protein by our method and by Ming–Brüschweiler's method are plotted in Figure 1. In the particular case of Cam, where data for different conformers are available, we observe significant variability in side-chain dynamics among these conformers (Supplementary Table S14). Results show that our neural network is sensitive to the minor structural changes between conformers, and is able to predict differences in methyl side chain dynamics (see examples).

Correlation coefficients for troponin C and fyn are remarkably high (correlation coefficients $r > 0.8$), whereas correlation is low ($r < 0.5$) for the heme protein. In this last case, the correlation coefficient is also low for Brüschweiler's predictions. This poor prediction for the heme protein can be due to the unusual rigidity exhibited by its side chains, possibly related to the interaction of this protein with the large heme prosthetic group (Flynn *et al.*, 2001).

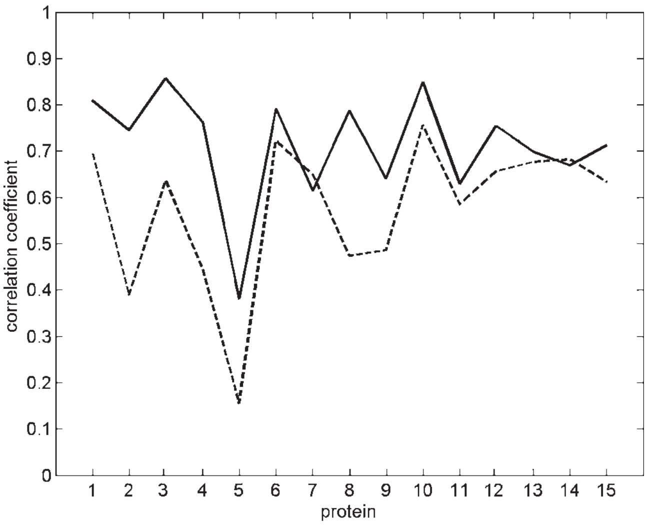

Fig. 1. Correlation coefficients between experimental and predicted S^2-values from the neural network predictor (solid line) and Ming–Brüschweiler's method (dotted line) for the proteins in the dataset: ^{1}ubq, ^{2}sh3, ^{3}trpC, ^{4}ALBP, ^{5}heme, ^{6}helix, ^{7}flav, ^{8}Tnfn, ^{9}Fnfn ^{10}fyn, ^{11}barn, ^{12}pin1, ^{13}CaM, ^{14}DHFR, ^{15}pdz (see Supplementary Tables S1, S10 and S11 for details).

3.3.3 Predictions for a new protein Predictions for the small protein eglin C (Clarkson *et al.*, 2006), which has not been included in the training set, revealed a correlation between predicted and experimentally determined methyl order parameters of 0.77 (Supplementary Fig. S5), supporting the predictive power of our neural network.

3.3.4 Comparison with molecular dynamics predictions Our neural network predictions of methyl side-chain S^2-values were compared with those computed from molecular dynamics simulations by several authors, for five proteins of the dataset; namely ubiquitin, TNfn3 and FNfn10 domains, barnase and Cam in complex with smMCLK (see Supplementary Table S13 for details). We observed an overall improvement in our predictions respect to molecular dynamics predictions, especially for the cases where low correlation was obtained by molecular dynamics (Table 3 and Supplementary Table S15). This result suggests that, in general, our neural network can be advantageously used as a reliable predictor of side-chain dynamics compared with more computationally demanding methods such as molecular dynamics.

3.4 Examples

Barnase and CaM are two examples of proteins in the dataset with available experimental information on dynamics for more than one conformational state. In this section, we compare experimental and predicted S^2-values for each methyl side-chain group in these two proteins in the free and bound states. Furthermore, we probe the predictive power of our neural network for detecting changes in side-chain dynamics upon changes in functional states.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

Table 3. Performance comparison between experimental and predicted S^2-values by molecular dynamics and the neural network

Protein	MD method	Solvent	MD	Neural network
ubiquitin[a]	AMBER99SB	Explicit	0.81	0.81
TNfn3[b]	CHARMM 22	Explicit	0.62	0.79
FNfn10[b]	CHARMM 22	Explicit	0.51	0.64
Barnase[c]	OPLS-AA/L	TIP4P	0.55	0.64
Cam[d]	FDPB	ZAP	0.73	0.71

Performance estimated by correlation coefficients in [a]Showalter *et al.*, 2007, [b]Best *et al.*, 2006 (read from the plot), [c]Zhuravleva *et al.*, 2007; average absolute normalized error in [d]Prabhu *et al.*, 2004.

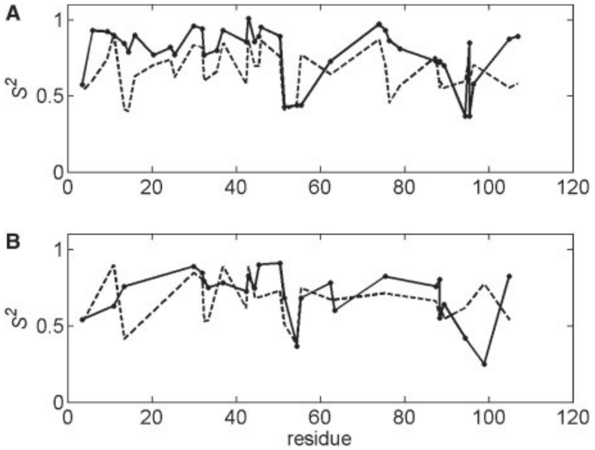

Fig. 2. S^2 experimental values (solid lines) and neural network predictions (dotted lines) of methyl side-chain for (**A**) free barnase (pdb 1a2p; A, $r = 0.63$) and (**B**) barnase complexed with barstar (pdb 1brs; A, $r = 0.64$).

3.4.1 Predictions for barnase Side-chain methyl dynamics for free barnase and in complex with barstar has been experimentally measured by ^{2}H relaxation (Zhuravleva *et al.*, 2007). Experimental and predicted S^2-values for the protein in the free and complexed states are shown in Figure 2. The correlation between experimental and predicted values is around $r = 0.64$ for both conformational states (P-values $= 6.6 \times 10^{-6}$ and 1.5×10^{-5}, respectively), while $r = 0.55$ was obtained by molecular dynamics in (Zhuravleva *et al.*, 2007).

We observed that residues val3 and ile55 were correctly predicted as residues containing two of the most flexible methyl groups in both states, as well as ile88 in the complexed state. Similarly, ala43 and leu33 in free barnase, and leu63 in barnase complexed with barstar were correctly identified among the most rigid side chains.

Table 4 lists the top 10 residues in barnase exhibiting significant changes ΔS^2 in its experimental and predicted methyl side-chain dynamics upon binding. The neural network was able to correctly identify eight of the 10 experimentally significant residues. Remarkably, residue ile51 is at the top of the list for both experimental and predicted ΔS^2. This residue, which is located in a region far from the binding interface (Fig. 3) and is involved in long-range communications, displays a significant change in dynamics (rigidification) upon barstar binding that cannot be explained based solely on the packing, as it has been noted in (Zhuravleva *et al.*, 2007). Furthermore, ile88, which appears second in the ranking for the predictions, has been also identified as a residue with changes in dynamics that cannot be explained by structural factors.

Table 4. Ranking of top 10 barnase residues with ‘largest’ changes in methyl side-chain dynamics upon binding with barstar

Type	Residue (ΔS^2)
Exp.	I51 (0.25); I55 (0.24); A46 (0.17); T107 (0.12); I88 (0.11)
Exp.	V10 (0.09); A74 (0.08); L63 (0.05); I76 (0.04); L33 (0.03)
Pred	**I51** (0.10); **I88** (-0.09); **T107** (0.07); **L33** (-0.07); **L42** (0.04)
Pred.	**I55** (-0.3); L20 (-0.02); **A46** (-0.01); **L63** (0.01); L14 (0.01)

Predictions corroborated by experimental observations are indicated in bold.

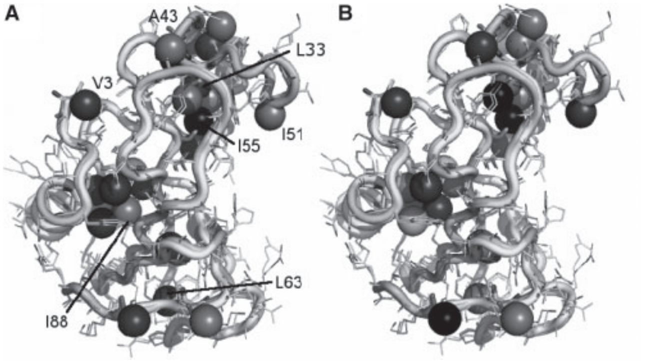

Fig. 3. (**A**) Experimental and (**B**) predicted changes in order parameter values ΔS^2 for barnase upon binding to barstar represented on the structural alignment of both states. Methyl groups are depicted by spheres. Darker gray spheres represent rigidification, whereas lighter gray spheres represent increase in dynamics upon binding. This image was created by using PyMOL (DeLano, 2002).

3.4.2 Predictions for Cam We compared experimental and predicted S^2-values for Cam (Lee *et al.*, 2000) in one of the closed free Ca^{2+}-bound conformations and in complex with the smooth muscle myosin light chain kinase (smMLCK). Experimental and predicted values of S^2 for each methyl group in both protein states are plotted in Figure 4. Correlation coefficients between experimental and predicted values in the free and bound states were found to be $r = 0.66$ (P-value $= 1.0 \times 10^{-7}$) and $r = 0.72$ (P-value $= 2.2 \times 10^{-12}$), respectively, while $r \sim 0.60$ was achieved by molecular dynamics in (Prabhu *et al.*, 2004) for the latter.

As it has been reported in other studies (Lee *et al.*, 2000), plasticity of ‘methionine puddles’ in CaM contribute to its broad target specificity, thus playing a key role in Cam binding. In fact, our method predicted a significant rigidification of residues met124, met109 and met71 (Fig. 5) upon binding, which have been recognized as important for activation of smMCLK (Lee *et al.*, 2000). Moreover, residue met76 is predicted to undergo a significant increase in flexibility upon binding, although this prediction could not be verified due to the lack of experimental information.

The 10 most significant experimental and predicted changes in S^2-values are listed in Table 5. Interestingly, in addition to the aforementioned four methionines, three other residues with

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

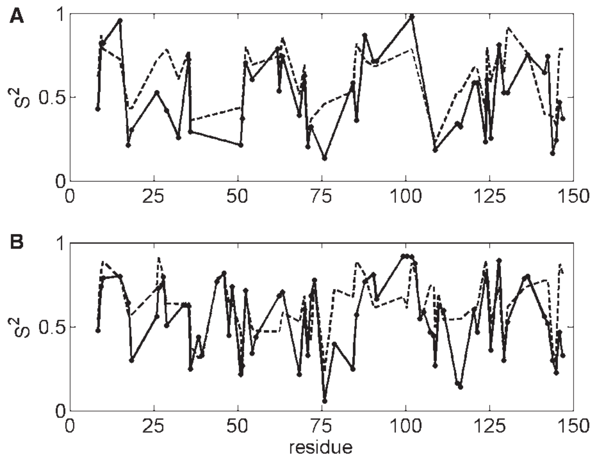

Fig. 4. S^2 experimental values (solid lines) and neural network predictions (dotted lines) of methyl side-chain for Cam (**A**) in the free state (pdb 1prw; A, $r = 0.66$) and (**B**) in complex with smMCLK (pdb 2o5g; A, $r = 0.72$).

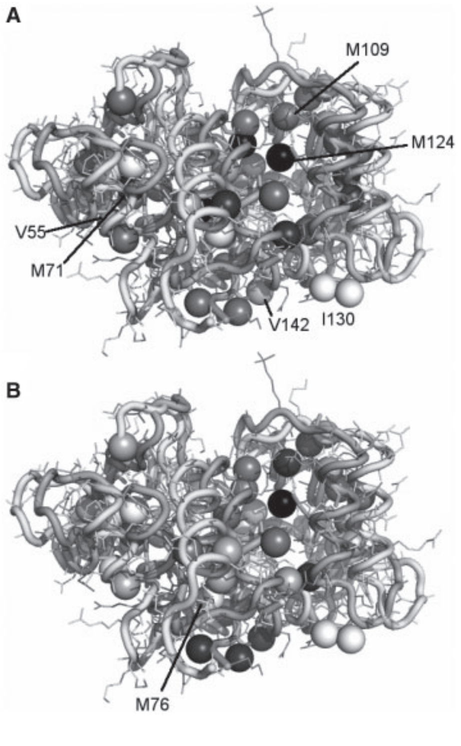

Fig. 5. (**A**) Experimental and (**B**) predicted changes in order parameter values ΔS^2-values for Cam upon smMCLK binding represented on the structural alignment of both states. Methyl groups are depicted by spheres. Darker gray spheres represent rigidification, whereas lighter gray spheres represent increase in dynamics upon binding. This image was created by using PyMOL (DeLano, 2002).

Table 5. Ranking of top 10 calmodulin residues with the 'largest' changes in methyl side-chain dynamics upon binding to smMLCK

Type	Residue (ΔS^2)
Exp.	M124 (0.58); L18 (0.43); I85 (-0.34); M72 (0.34); V55 (-0.27)
Exp.	I130 (-0.23); V142 (-0.23); L116 (-0.21); L69 (-0.18); A15 (-0.16)
Pred	**M124** (0.48); **V142** (0.38); **I130** (-0.36); **M109** (0.26); I63 (-0.26)
Pred.	**M76** (-0.22); **V55** (-0.21); I52 (-0.21); **M71** (0.17); **I85** (0.14)

Predictions corroborated by experimental observations are indicated in bold.

significant experimental changes in dynamics ΔS^2, val142, ile130 and val55, were also identified by our predictions. These results are remarkable, given the fact that only 65% of methyl S^2-values in free Cam were sufficiently resolved by the NMR experimental characterization (Lee *et al.*, 2000), and therefore proper values of differences ΔS^2 could be only specified for some of the CaM methyl groups in the dataset.

4 DISCUSSION

Protein dynamics is believed to influence biological function through diverse mechanisms, which are still poorly understood. Thus, prediction of protein conformational flexibility from sequence or structural information has a considerable importance in understanding protein function, and is becoming relevant in fields such as protein modeling and engineering. Here, we focus on the assessment of sub-nanosecond dynamics of protein side chains, which is more diverse than backbone dynamics, and has been shown to be fundamentally linked to processes such as protein–ligand and protein–protein associations, and in particular to the transmission of allosteric communications. Using a dataset of 18 proteins, we trained an artificial neural network for the prediction of methyl side-chain generalized order parameters from the 3D structures of proteins. The network parameterization contains 10 input parameters, which are classified in two groups: Geometry-based parameters and knowledge-based potentials. These input parameters considered side-chain and backbone structural characteristics. Validation of the neural network predictions yielded to an average correlation coefficient between experimental and predicted values of the generalized order parameter $r = 0.71 \pm 0.029$. Furthermore, our results indicate that the input parameters, which are more strongly correlated to the experimental values of the generalized order parameter, are the methyl carbon packing density, its distance to the C_α atom, and the knowledge-based pair-wise contact potential between the methyl carbon and neighboring residues. Some of these characteristics were previously taken into account in a simple model proposed by Ming and Brüschweiler for the prediction of methyl side-chain order parameters. The inclusion of additional characteristics in the neural network leads to a greater predictive power of our method in comparison with their model. However, the lack of strong correlation between each of these additional input parameters with the experimental methyl side-chain order parameters suggests that other structural characteristics remain to be found and incorporated in the neural network. It is worth noting that we used a modified version of the Ming and Brüscheweiler's methyl carbon packing density, which also considers the packing density of the methyl carbon's neighboring atoms. This correction was shown to lead to overall improved predictions, perhaps due to the fact that in a way it takes into account the collective re-orientational dynamics of side chains. In addition, our method, which is simple and easy to implement, was compared with molecular dynamics simulations for the prediction of methyl order parameters in five examples of

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014

proteins (ubiquitin, TNfn3 domain and FNfn10 domain, barnase and Cam). Based on the analyzed examples, the neural network shows an overall improvement in prediction performance respect to molecular dynamics, especially for the cases where low correlation was obtained by molecular dynamics. This result suggests that, in general, our neural network can be advantageously used as a robust predictor of side-chain dynamics, particularly in large proteins where molecular dynamics simulations might be computationally demanding.

The sensitivity of our predictions to details of the protein structure was illustrated with the examples of two proteins (Cam and barnase), exhibiting changes in side-chain dynamics with no significant structural changes upon ligand binding. Results show that the predicted changes in methyl side-chain order parameters upon ligand binding are in good agreement with experimental data. In particular, methionine methyl groups in Cam, which are known to experience a wide range of changes in methyl order parameters upon ligand binding (Igumenova *et al.*, 2006) are generally correctly predicted. In these examples, several residues with predicted significant changes in methyl order parameters have been experimentally identified as amino acids involved in long-range interactions or protein–ligand binding. Thus, considering the importance of side-chain dynamics in binding and allostery, these results suggest that our neural network model may assist in understanding these biological processes and in the design of functional proteins. In addition, we believe that our method can be combined with other approaches which predict backbone dynamics and protein motions on different time scales in order to have a more complete description of protein dynamics and to elucidate its role in protein function. Integration of these different types of motion in a single model remains a challenge.

Conflict of Interest: none declared.

REFERENCES

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Best,R.B. *et al.* (2005) What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. *J. Mol. Biol.*, **349**, 185–203.

Best,R.B. *et al.* (2006) Relation between native ensembles and experimental structures of proteins. *Proc. Natl Acad. Sci. USA*, **103**, 10901–10906.

Boyer,J.A. and Lee,A.L. (2008) Monitoring aromatic picosecond to nanosecond dynamics in protein via ^{13}C relaxation: expanding perturbation mapping of the rigidifying core mutations, V54A, in Eglin c. *Biochemistry*, **47**, 4876–4886.

Clarkson,M.W. *et al.* (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein'. *Biochemistry*, **45**, 7693–7699.

DeLano,W. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific , Palo Alto, CA.

Eldon,L. *et al.* (2007) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D407.

Flynn,P.F. *et al.* (2001) Main chain and side chain dynamics of a heme protein: 15N and 2H NMR relaxation studies of R. capsulatus ferrocytochrome c2. *Biochemistry*, **40**, 6559–6569.

Frederick,K.K. *et al.* (2007) Conformational entropy in molecular recognition by proteins. *Nature*, **448**, 325–330.

Friedland,G.D. *et al.* (2008) A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.*, **380**, 757–774.

Fuentes,E.J. *et al.* (2004) Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J. Mol. Biol.*, **335**, 1105–1115.

Goodey,N.M. and Benkovic,S.J. (2008) Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.*, **4**, 474–482.

Haykin,S. (1998) *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Englewood Cliffs, NJ.

Henzler-Wildman,K. and Kern,D. (2007) Dynamic personalities of proteins. *Nature*, **450**, 964–972.

Hu,H. *et al.* (2005) Relating side-chain mobility in proteins to rotameric transitions: insights from molecular dynamics simulations and NMR. *J. Biomol. NMR*, **32**, 151–162.

Igumenova,T.I. *et al.* (2006) Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.*, **106**, 1672–1699.

Jones,W.C. *et al.* (1976) Nuclear magnetic resonance studies of sperm whale myoglobin specifically enriched with 13C in the methionine methyl groups. *J. Biol. Chem.*, **251**, 7452–7460.

Kabsch,W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Lee,A. and Wand,A.J. (2001) Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature*, **411**, 501–504.

Lee,A. *et al.* (2000) Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat. Struct. Biol.*, **7**, 72–77.

Lipari,G. and Szabo,A. (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, **104**, 4546–4559.

Lovell,S.C. *et al.* (2000) The penultimate rotamer library. *Proteins: Struct. Funct. Genet.*, **40**, 389–408.

Ming,D. and Brüschweiler,R. (2004) Prediction of methyl-side chain dynamics in proteins. *J. Biomol. NMR*, **29**, 363–368.

Mittermaier,A. *et al.* (1999) Analysis of deuterium relaxation-derived methyl axis order parameters and correlation with local structures. *J. Biomol. NMR*, **13**, 181–185.

Mittermaier,A. *et al.* (2003) Correlation between ^2H NMR side-chain order parameters and sequence conservation in globular proteins. *J. Am. Chem. Soc.*, **125**, 9004–9005.

Namanja,A.T. *et al.* (2007) Substrate recognition reduces side-chain flexibility for conserved hydrophobic residues in human Pin1. *Structure*, **15**, 313–327.

Nicholson,L.K. *et al.* (1992) Dynamics of methyl groups in proteins as studied by proton-detected 13C NMR spectroscopy. Application to the leucine residues of staphylococcal nuclease. *Biochemistry*, **31**, 5253–5263.

Nissen,S. (2003) Implementation of a Fast Ariticial Neural Network Library (fann). *The University Report*, Departament of Computer Science, University of Copenhagen.

Prabhu,N.V. *et al.* (2003) Dynamics and entropy of a calmodulin-peptide complex studied by NMR and molecular dynamics. *Biochemistry*, **42**, 562–570.

Prabhu,N.V. *et al.* (2004) Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method. *J. Comput. Chem.*, **5**, 2049–2064.

Seavey,B.R. *et al.* (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.

Showalter,S.A. *et al.* (2007) Toward quantitative interpretation of methyl side-chain dynamics from NMR by molecular dynamics simulations. *J. Am. Chem. Soc.*, **129**, 14146–14147.

Trbovic,N. *et al.* (2009) Protein side-chain dynamics and residual conformational entropy. *J. Am. Chem. Soc.*, **131**, 615–622.

Trott,O. *et al.* (2008) Protein conformational flexibility prediction using machine learning. *J. Magn. Reson.*; **192**, 37–47.

Wintrode,P.L. and Privalov,P.L. (1997) Energetics of target peptide recognition by calmodulin: a calorimetric study. *J. Mol. Biol.*, **266**, 1050–1062.

Yang,D. and Kay,L.E. (1996) Contribution to conformational entropy arising from bond-vector fluctuations measured from NMR-derived order parameters: application to protein folding. *J. Mol. Biol.*, **263**, 369–382.

Zhuravleva,A *et al.* (2007) Propagation of dynamic changes in barnase upon binding of barstar: an NMR and computational study. *J. Mol. Biol.*, **367**, 1079–1092.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Luxembourg on April 28, 2014