Methyl side-chain dynamics prediction based on protein structure

Pablo Carbonell1,2 and Antonio del Sol1,*

1Fujirebio Inc., Bioinformatics Research Unit, Research and Development Division, Komiyacho, Hachioji-shi, Tokyo 192-0031, Japan and 2University 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 (S2) 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 ± 0.029. Further analysis revealed that the order parameter depends more strongly on the methyl carbon distance to the 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 S2 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: antdelso@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Protein 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 (Goeddey 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 (Goeddey 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 in 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; Tribovic 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 time, which explores side-chain motions in the picosecond-to-nanosecond time regime, can be further interpreted in the form of Lipari-Szabo S2 order parameters and τe internal correlation times.

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
A different approach for modeling side-chain dynamics considered TNfn3 domain, FNfn10 domain, barnase and Cam) were, in general, 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 $C\alpha$ atom and the knowledge-based pairwise 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, pairwise 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).

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

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 definition, we used the following four packing density parameters:

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Methyl group</th>
<th>$S^2$-values</th>
<th>$S^2$-values (error $&lt; 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>CB</td>
<td>324</td>
<td>260</td>
</tr>
<tr>
<td>ILE</td>
<td>CD1</td>
<td>320</td>
<td>309</td>
</tr>
<tr>
<td>ILE</td>
<td>CG2</td>
<td>253</td>
<td>233</td>
</tr>
<tr>
<td>LEU</td>
<td>CD1</td>
<td>380</td>
<td>342</td>
</tr>
<tr>
<td>LEU</td>
<td>CD2</td>
<td>351</td>
<td>320</td>
</tr>
<tr>
<td>MET</td>
<td>CE</td>
<td>179</td>
<td>177</td>
</tr>
<tr>
<td>THR</td>
<td>CG2</td>
<td>248</td>
<td>225</td>
</tr>
<tr>
<td>VAL</td>
<td>CG1</td>
<td>322</td>
<td>310</td>
</tr>
<tr>
<td>VAL</td>
<td>CG2</td>
<td>320</td>
<td>312</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2697</td>
<td>2488</td>
</tr>
</tbody>
</table>

$S^2$-values for each type of amino acid and methyl group in the dataset.
expression, we define the packing density of methyl carbon $i$ as:

$$P_i = \sum_j C_{ij} C_{ej}$$  \hfill (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:

$$\tilde{K}_i = \sum_{j=1}^{N_b} P_j$$

where $j$ goes through the number of consecutive dihedral angles from the backbone and $N_b$ 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 $C_b$:

$$E_i = r_{C_i}$$  \hfill (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 $\chi$ 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:

$$\Delta \chi = \chi - \chi_0$$  \hfill (5)

where $\chi_0$ 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 $j$ type defined as:

$$K_{ij} = -\sum_{r} \log(f_{ij}(r))$$

$$f_{ij}(r) = \frac{1}{e^{r/\sigma_{ij}} - 1}$$  \hfill (6)

where $r$ is the distance between the carbon in the methyl group $i$ and the closest heavy atom in the contacting $j$ amino acid, and $\sigma_{ij}$ is the frequency of occurrence of such contact for this type of amino acid and distance $r$, 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 $\varepsilon < 0.005$, 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). Furthermore, 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 $C_n$), pair-wise contact potential, and methyl packing density.
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 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–Brüschweiler’s prediction method (Ming and Bruchweiler, 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).

#### 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, TNF$\alpha$3 and FN$\alpha$10 domains, barnase and Cam in complex with smMCLK (see Supplementary Table S13 for details). We observed an overall improvement in our predictions with 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.
Table 3. Performance comparison between experimental and predicted changes in side-chain dynamics upon binding.

<table>
<thead>
<tr>
<th>Protein</th>
<th>MD method</th>
<th>Solvent</th>
<th>MD</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>ubiquitin</td>
<td>AMBER99SB</td>
<td>Explicit</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>TNFαB</td>
<td>CHARMm22</td>
<td>Explicit</td>
<td>0.62</td>
<td>0.79</td>
</tr>
<tr>
<td>FNIII</td>
<td>CHARMm22</td>
<td>Explicit</td>
<td>0.51</td>
<td>0.64</td>
</tr>
<tr>
<td>Barnase</td>
<td>OPLS-AA-L</td>
<td>TIP4P</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td>CamD</td>
<td>FDPB</td>
<td>ZAP</td>
<td>0.73</td>
<td>0.71</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Residue</th>
<th>ΔS&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp.</td>
<td>I51</td>
<td>0.25</td>
</tr>
<tr>
<td>Exp.</td>
<td>I55</td>
<td>0.24</td>
</tr>
<tr>
<td>Exp.</td>
<td>A46</td>
<td>0.17</td>
</tr>
<tr>
<td>Exp.</td>
<td>T107</td>
<td>0.12</td>
</tr>
<tr>
<td>Exp.</td>
<td>I88</td>
<td>0.11</td>
</tr>
<tr>
<td>Pred</td>
<td>I51</td>
<td>0.10</td>
</tr>
<tr>
<td>Pred</td>
<td>I55</td>
<td>-0.09</td>
</tr>
<tr>
<td>Pred</td>
<td>T107</td>
<td>0.07</td>
</tr>
<tr>
<td>Pred</td>
<td>L33</td>
<td>-0.07</td>
</tr>
<tr>
<td>Pred</td>
<td>L42</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Predictions corroborated by experimental observations are indicated in bold.

Fig. 3. (A) Experimental and (B) predicted changes in order parameter values ΔS<sup>2</sup> 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).

(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.

3.4.2 Predictions for Cam

We compared experimental and predicted ΔS<sup>2</sup>-values for Cam (Lee et al., 2000) in one of the closed free Ca<sup>2+</sup>-bound conformations and in complex with the smooth muscle myosin light chain kinase (smMLCK). Experimental and predicted values of ΔS<sup>2</sup> 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 × 10<sup>-7</sup>) and r = 0.72 (P-value = 2.2 × 10<sup>-12</sup>), respectively, while r ≈ 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 smMLCK (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<sup>2</sup>-values are listed in Table 5. Interestingly, in addition to the aforementioned four methionines, three other residues with
Methyl side-chain dynamics

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 $\Delta S^2$-values for Cam upon smMLCK 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).

significant experimental changes in dynamics $\Delta S^2$, 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

<table>
<thead>
<tr>
<th>Type</th>
<th>Residue ($\Delta S^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp.</td>
<td>M124 (0.58); L18 (0.43); I85 (-0.34); M72 (0.34); V55 (-0.27)</td>
</tr>
<tr>
<td>Exp.</td>
<td>I130 (-0.23); V142 (-0.23); L116 (-0.21); L69 (-0.18); A15 (-0.16)</td>
</tr>
<tr>
<td>Pred</td>
<td>M124 (0.48); V142 (0.38); I130 (-0.36); M109 (0.26); I63 (-0.26)</td>
</tr>
<tr>
<td>Pred</td>
<td>M76 (-0.22); V55 (-0.21); I52 (-0.21); M71 (0.17); I85 (0.14)</td>
</tr>
</tbody>
</table>

Predictions corroborated by experimental observations are indicated in bold.

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

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 $C_\alpha$ 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 Brüschweiler’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
proteins (ubiquitin, TNF3 domain and FNF10 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 may 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 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 advantageously used as a robust tool in molecular dynamics, especially for the cases where low correlation studies have been experimentally identified with the examples of two proteins (Cam and barnase).

**REFERENCES**


Conflict of Interest: none declared.