JCTC Journal of Chemical Theory and Computation

Article

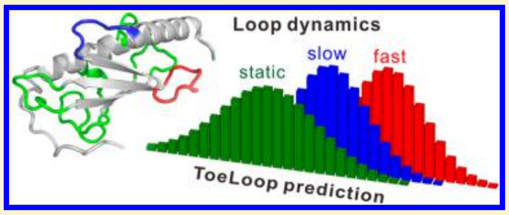
pubs.acs.org/JCTC

Decoding the Mobility and Time Scales of Protein Loops

Yina Gu,[†] Da-Wei Li,[‡] and Rafael Brüschweiler*[,†,‡]

[†]Department of Chemistry and Biochemistry and [‡]Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio 43210, United States

Supporting Information



ABSTRACT: The flexible nature of protein loops and the time scales of their dynamics are critical for many biologically important events at the molecular level, such as protein interaction and recognition processes. In order to obtain a predictive understanding of the dynamic properties of loops, 500 ns molecular dynamics (MD) computer simulations of 38 different proteins were performed and validated using NMR chemical shifts. A total of 169 loops were analyzed and classified into three types, namely fast loops with correlation times <10 ns, slow loops with correlation times between 10 and 500 ns, and loops that are static over the course of the whole trajectory. Chemical and biophysical loop descriptors, such as amino-acid sequence, average 3D structure, charge distribution, hydrophobicity, and local contacts were used to develop and parametrize the ToeLoop algorithm for the prediction of the flexibility and motional time scale of every protein loop, which is also implemented as a public Web server (http://spin.ccic.ohio-state.edu/index.php/loop). The results demonstrate that loop dynamics with their time scales can be predicted rapidly with reasonable accuracy, which will allow the screening of average protein structures to help better understand the various roles loops can play in the context of protein−protein interactions and binding.

1. INTRODUCTION

Protein loops play an important role for many protein properties and function. A primary purpose of loops is to connect secondary structural elements, i.e. α-helices and β-strands, allowing proteins to adopt a globular shape or to serve as linkers between consecutive domains in multidomain proteins. Loop lengths often exceed what is necessary to merely serve as connectors allowing them to play additional roles, such as the specific recognition and binding of proteins, which can be of high functional importance.[1] Different loops display a variable level of structural plasticity, which covers fully disordered behavior, limited flexibility, and well-defined structure. The details of loop plasticity and the associated motional time scales are key elements for their biological function. Fundamental questions surrounding current frameworks of molecular recognition, such as induced fit, conformational selection, and population shift, require a proper understanding of both the conformational space and the time scales sampled by flexible loop regions that are involved in binding.[2]

Loop motions are known to occur on a wide range of time scales from picoseconds to milliseconds. With advances in computer power, molecular dynamics (MD) simulations[3] have been used as a powerful tool to characterize and visualize protein dynamics at atomic detail on the time scales from picosecond to microsecond.[4] Coarse-grained descriptions of protein dynamics using elastic network models have been developed for the prediction of protein dynamics from average 3D structures.[5] Other approaches have been developed to predict protein dynamics and disorder based on protein sequences,[6−13] local densities,[14] atomic contacts,[15,16] and structural prototypes,[17] or other physical-chemical parameters, including chemical shifts.[18] With the exception of MD, these approaches focus on classifying dynamics according to the magnitude of dynamics fluctuations independent of the underlying time scales. MD simulations can in addition provide detailed time scale information depending on the lengths and quality of the trajectories.

In this work, we use MD simulations to obtain an increasingly predictive understanding of the time scales of protein loop dynamics. We first categorize loops of 38 proteins according to their overall flexibility. Second, by identifying the determining factors of loop dynamics, including loop length, amino-acid composition, hydrogen-bonding patterns, and atomic contact sums, we construct a model that rapidly and reliably predicts protein loop time scales. This new computational tool, which is termed “Time scales of every Loop” or ToeLoop and which is implemented as a Web server, allows one to characterize loop dynamics from average structures and elucidate their functional role, for example, during the course of a molecular recognition event.

2. MATERIALS AND METHODS

2.1. Molecular Dynamics Simulations and Analysis.

For the parametrization of ToeLoop, a total of 38 proteins were selected based on the availability of medium-to-high resolution X-ray crystal structures in the protein databank (PDB) and the availability of NMR chemical shifts of proteins in their native, monomeric state in the BMRB database. The PDB and BMRB codes of all proteins used in this work are listed in the Supporting Information. The X-ray crystal structures were used

Received: December 2, 2014
Published: February 2, 2015

DOI: 10.1021/ct501085y

as starting points of MD simulations performed with the ff99SB_$\varphi\psi$(g24;CS) force field[19] using the Gromacs 4 package.[20−23] The proteins were solvated in explicit water in a cubic box using the TIP3P water model fixed by the SETTLE algorithm.[24] Each protein was adjusted to neutral pH by adjusting its protonation state, and each protein system had a net charge of zero by the addition of counterions. Long-range electrostatic interactions were treated using the PME algorithm with 1.2 Å spacing at a 8 Å cutoff, and van der Waals interactions had a cutoff distance of 10 Å. All bond lengths involving hydrogen atoms were constrained using the LINCS algorithm.[25] After application of a standard energy-minimization and equilibration protocol, each system was run for a 500 ns long production MD with a 2 fs time step at constant 300 K temperature and 1 atm pressure (NPT). 5000 snapshots were stored for analysis with a sampling rate of one frame every 100 ps. The present work required approximately 300,000 core hours for a cumulative total of 19 μs of MD trajectories.

Protein secondary structures and solvent accessibilities were analyzed by the DSSP program[26] based on the initial crystal structures. The isotropic reorientational eigenmode dynamics (iRED) method[27,28] was applied to the MD trajectories to determine the NMR S^2 order parameters[29] of the Cα−Cβ bonds of all residues (except for glycine). The Cα−Cβ bond vectors, which show an overall similar dynamics behavior as backbone N−H vectors, were chosen as representative probes of the residue they belong to in order to assess both the amount of motion and the dominant time scales. The trajectory time windows used for iRED were either the full 500 ns trajectory or nonoverlapping blocks (windows) of 10 ns each. In the case of a 10 ns window length, the order parameters were calculated for each window and then averaged over all 50 windows.

2.2. Chemical Shift Prediction from MD Ensembles. To validate the quality of each MD trajectory, we compared the experimental NMR chemical shifts with the predicted chemical shifts using both our recently developed PPM[30] software as well as the program SHIFTS.[31] Most of the structure-based chemical shifts prediction programs, including SPARTA+, SHIFTX, and CAMSHIFT, are knowledge-based and are fitted directly against the experimental static protein structures. Thus, some ensemble-averaging effects are implicitly included in their prediction, which makes these programs unsuitable for the assessment of the quality of conformational ensembles. In fact, when provided with a Boltzmann ensemble and averaged explicitly, the predictions are often less accurate than for an average structure.[32,33] PPM was specifically parametrized using conformational ensembles from long MD simulations of proteins, which makes it suitable for the quantitative assessment of MD trajectories.[30] SHIFTS was originally based on approximate analytical relationships between chemical shifts and atomic coordinates derived from quantum-chemical calculations. Consistent with the fact that the experimental chemical shift of a given nucleus reflects the Boltzmann-weighted average of the ’instantaneous’ chemical shift over a large number of conformational substates, it was previously found that the chemical shift prediction generally improves when longer MD and better sampled trajectories are used.[34−36] To ensure that the generated MD ensembles are physically reasonable, the chemical shifts of all 38 proteins used in this study were predicted using PPM and SHIFTS by averaging over time windows of increasing lengths always starting from the beginning of the trajectory up to 500 ns. The prediction error was expressed by the chemical shift root-mean-square difference (RMSD, in units of ppm).

2.3. Prediction Parameters and Algorithm of ToeLoop Predictor. A flowchart of the ToeLoop prediction approach is depicted in Figure 1 containing two main stages. In the first part, loops are selected as regions that neither adopt an α-helix (H) nor an extended strand (E) according to the DSSP program. Then all loops are sorted into “static” and “dynamic” loops, while in the second part the dynamic loops are subdivided into “fast” and “slow” loops with correlation times smaller or larger than 10 ns, respectively. In the first stage, after reading in the PDB file of the corresponding crystal structure, a loop matrix is constructed containing for each loop 13 parameters, which are described below. Based on the MD simulations, each loop is assigned either a score of 1, 0, −1 depending on whether the loop undergoes fast dynamics, slow dynamics, or whether it is static, respectively. The training algorithm is used to parametrize a numerical model by optimizing the relative weights of the 13 loop parameters to reproduce the dynamics scores (score I) using linear least-squares fitting based on singular value decomposition (SVD). Trained on a total of 169 loops from 38 proteins, the best predictor (with the minimum χ^2 error) between predicted scores and true scores is

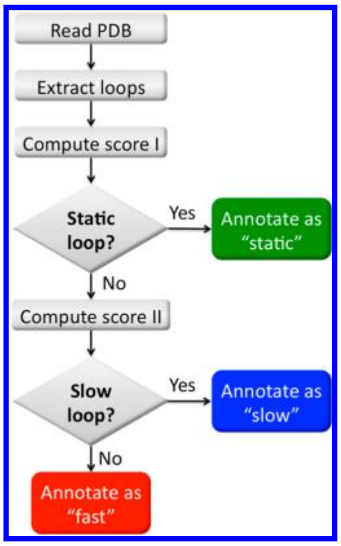


Figure 1. Flowchart describing the loop dynamics prediction protocol ToeLoop.

$$\text{score I} = \sum_{i=1}^{5} a_i f(R_i) + \sum_{j=1}^{4} b_j f(S_j) + \left(9.5\langle V\rangle - 38.9\langle H\rangle + 54.3L - \sum_{i}\sum_{j\neq i} C_{ij}\right) \times 10^{-3} - 103.66 \tag{1}$$

The different parts of eq 1 incorporate a total of 13 descriptors, which are the relative amino-acid frequencies $f(R_i)$ $(0 \le f(R_i) \le 1)$ of histidines (R_1), arginines (R_2), lysines (R_3), all negatively charged residues (R_4), and all neutral residues (R_5) with prefactors $a_1,...,a_5$ = (102.3, 101.8, 102.5, 102.8, 103.0) for each loop. For instance, for a loop consisting of ten residues containing two aspartic acids and one glutamic acid, the relative

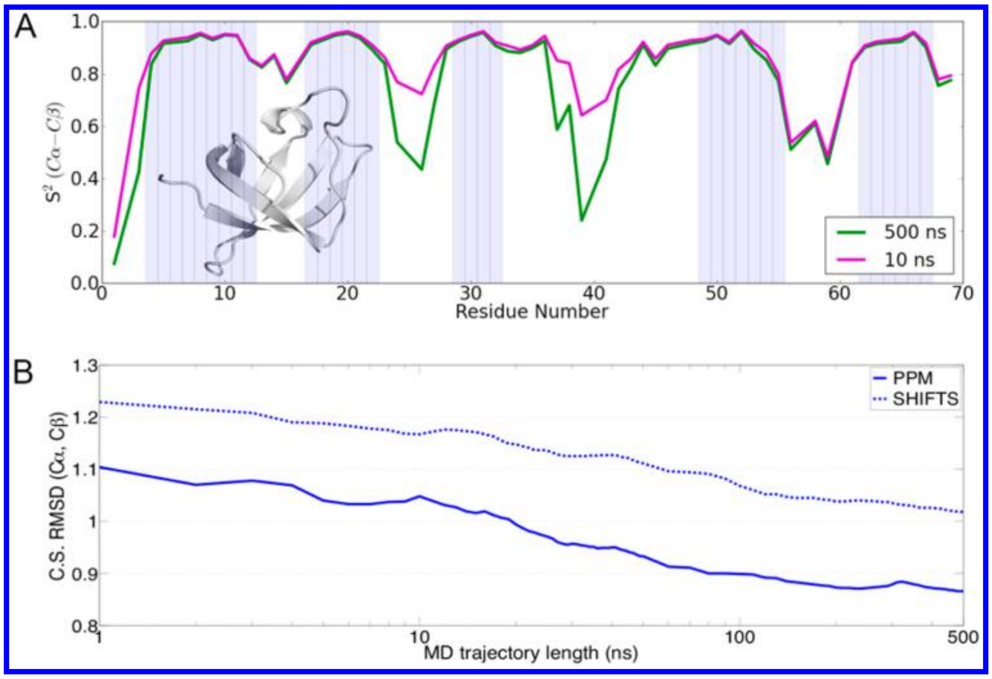


Figure 2. Mobility and time scales of protein loops exemplified for protein Csp A (PDB 1MJC). (A) $C\alpha-C\beta$ bond vector order parameters S^2 calculated from the same 500 ns MD trajectory (green) and an average over 50 consecutive 10 ns MD blocks (magenta). The colored bars indicate the protein secondary structural elements: β-strands (blue) and loops (white). (B) Average $C\alpha$ and $C\beta$ chemical shift (C.S.) RMSD between experimental chemical shifts and chemical shifts (in units of ppm) predicted by the programs SHIFTS and PPM for increasing MD trajectory length from zero to 500 ns.

frequency of negatively charged residues $f(R_4)$ is 0.3. Next, using the secondary structure identification software DSSP,[26] the frequencies of four different loop secondary structures $f(S_j)$ were determined, namely isolated β-bridges (S_1), 3_{10} helices (S_2), hydrogen bonded turns (S_3), and bends (S_4) with prefactors $b_1,...,b_4$ = (−1.09, −0.05, 0.20, 0.67). The contact sum C_{ij} of each heavy atom i in a loop is determined as the sum of interactions of atom i with all other heavy atoms j in the protein that are not part of the same loop, where $C_{ij} = \exp(-r_{ij}/r_0)$, r_{ij} is the distance between atom i and j, and r_0 is an effective interaction distance set to 3.4 Å.[16] Furthermore, the loop length L corresponding to the number of amino acids of the loop, the average solvent accessibility surface area per loop residue $\langle V\rangle$ provided by DSSP, and the average hydrophobicity index $\langle H\rangle$ over all residues in a loop are used as additional parameters. Application of the model shows that a loop score I < −0.33 yields an optimal prediction accuracy of 81% for static loops. By definition, Accuracy = (TP + TN)/N, where TP is the number of loops with a true positive outcome, TN is the number of loops with a true negative outcome, and N is the total number of loops.

In the second stage, those loops classified as being dynamic are subdivided into fast and slow loops using a sequence-based scoring algorithm. Score II for each dynamic loop is the sum of relative frequencies over all 20 amino acids in a loop

$$\text{score II} = \sum_{k=1}^{20} a_k \cdot f(A_k) \tag{2}$$

where $f(A_k)$ is the relative frequency of amino-acid type A_k where $A_1,...,A_{20}$ = (R, N, Q, H, Y W, T, G, L, V, I, K, D, E, P, S, A, M, C, F) and prefactors $a_1,...,a_{20}$ = (1.08, 0.16, 0.43, 1.66, 3.29, 2.06, 1.10, 0.17, 0.63, 0.72, 4.29, −0.90, −0.44, −0.16, −2.26, −0.24, −0.09, −5.53, −0.63, −0.36). A 74% prediction accuracy could be achieved for a threshold value of score II of −0.98, whereby loops with predicted scores above the threshold are annotated by ToeLoop as "slow loops" and below the threshold as "fast loops". The ToeLoop predictor is written in the Python programming language, which powers a public ToeLoop Web server accessible at http://spin.ccic.ohio-state.edu/index.php/loop.

3. RESULTS AND DISCUSSION

3.1. Classification of Protein Loop Mobility and Time Scales from MD Trajectory. The overall stability of the 500 ns MD trajectories of all 38 proteins was assessed from the root-mean-square deviation (RMSD) of each snapshot relative to the X-ray crystal structure. The backbone RMSD averaged over all 38 MD trajectories is 2.3 ± 0.4 Å reflecting good overall stability of the MD simulations. Average chemical shifts were calculated for $C\alpha$, $C\beta$, C', N^H, and H^N nuclei from the MD ensembles for increasing window length. Averaging over longer trajectories generally lead to smaller, i.e. better, chemical shift RMSDs. This is consistent with the physical origin of NMR chemical shifts reporting on time and ensemble averages into the millisecond range, and it suggests that for the majority of the systems used here the MD ensembles become more realistic as the trajectory lengths increase. The improvement of the average $C\alpha$ and $C\beta$ chemical shift RMSD from 0 to 500 ns window length is shown in Figure 2B for protein Csp A (PDB: 1MJC) with identical trends observed both for SHIFTS and PPM programs. Similar trends of chemical shift RMSD are shown in Figure S2 for six additional proteins. Figure 2A depicts the S^2 order parameters of the $C\alpha-C\beta$ bond vectors determined from both the 10 and 500 ns window length, whereby for the 10 ns window length the calculated S^2 order parameters were averaged over 50 consecutive, nonoverlapping 10 ns windows. As can be seen in Figure 2A, the loops comprising residues 23−28 and 37−41 display significantly decreased S^2 values when going from the 10 ns to the 500 ns window length, which is evidence that both of these loops are flexible during the MD trajectory on a time scale between 10 and 500 ns. All loops with such behavior displaying correlation times slower than 10 ns but faster than 500 ns are classified as "slow". It is worth emphasizing that for the loops denoted as static, it cannot be excluded with certainty that they undergo conformational dynamics on the microsecond time scale or slower, which is beyond the time scales probed by our MD simulations. With further increase of computer power, it will be possible to examine even slower time scale dynamics that can be experimentally monitored, for example, by relaxation dispersion measurements and residual dipolar couplings. All β-strand regions and some of the other loops show only minor or no difference between the two window lengths for their chemical shift prediction RMSDs and order parameters. For example, the order parameters of loop residues 56−61 exhibit consistently low values (<0.6) for the 10 ns window without a further reduction during the 500 ns window. This indicates that internal motions predominantly occur on time scales faster than 10 ns. Consequently, this loop is categorized as "fast". For the loop residues 13−16, relatively high order parameters are found (on average >0.8), which suggest the absence of large amplitude internal motions over the duration of the trajectory and thus it is considered as "static".

All 38 proteins were analyzed using the same methodology used for protein Csp A, and the results of a representative set of six proteins are depicted in Figure S1. For the 38 proteins, a total of 169 loops were identified and inspected with respect to the specific motional behavior: each loop was assigned to one of the 3 categories "slow", "fast", or "static" based on the selection criteria given in Table 1. The assignment to a specific category is based on the average S^2 values over the 10 and 500 ns windows and their difference ΔS^2. Some loops display both fast and slow dynamics as is displayed in Figure S6. This is one of the reasons why fast and slow loops are not as easily discriminated with respect to each other when compared to the discrimination between static and mobile loops. For simplicity, all loops that are fast and slow are assigned to the "slow" category.

Table 1. Order Parameters Criteria for Loop Classification

loop category	subcategory	S^2 (10 ns)	S^2 (500 ns)	ΔS^2 (10−500 ns)
static		≥0.8	≥0.8	
fast		<0.8	<0.8	<0.2
slow	slow and fast	<0.8	<0.6	>0.2 and <0.4
	slow only		<0.6	≥0.4

This leads to a total of 58 slow loops and 50 fast loops from 28 and 29 proteins, respectively, that were included in this study. Although a larger number of static loops could be identified, a total of 61 static loops from 30 proteins were randomly selected to constitute around 1/3 of the total number of loops in the database, to achieve a balanced parametrization for all three categories. Another set of 63 static loops from 18 proteins that were not part of the training set was used as a validation set during the first stage of ToeLoop prediction. A leave-one-out validation strategy is applied to prevent overfitting in all parametrization procedures.

3.2. Analysis of Amino-Acid Composition and Other Prediction Parameters.

Statistical analysis of amino-acid frequencies in the three loop categories reveals characteristic patterns for the amino-acid composition for each type of loops. Figure 3 shows the relative frequencies of each amino-acid type either normalized relative to the amino-acid type (Figure 3B) or normalized relative to the three loop categories (Figure 3C). Figure 3B reveals key information about the probability distribution over the three dynamics categories for each amino-acid type offering a useful source of information for predicting loop behavior based on amino-acid frequencies. Certain amino acids, such as methionine, aspartate, and lysine, have a clear propensity for fast loops, while other amino acids have a preference for slow loops, which includes tryptophans, threonines, valines, and histidines. Other amino acids, such as cysteines, tyrosines, isoleucines, phenylalanines, and leucines, are often part of static loops (Figure 3B).

Interestingly, the three amino-acid types that are most characteristic for fast loops are ones that are either positively charged (Lys), negatively charged (Asp), or neutral (Met). Methionine is known to often play a key role in protein−protein interactions that require a high level of flexibility.[37,38] The other charged residues (Arg and Glu) do not have a preference for fast loops. In fact, Arg has a propensity for both slow and static loops, while Glu has a slight preference for static loops. It is important to note that this statistical analysis does not provide direct insights into the causality relationship between amino-acid type and loop time scales. For example, it is not *a priori* known whether methionines actively accelerate loop motions, since it is possible that evolution placed methionines in loops that are already intrinsically fast. On the other side of the spectrum, isoleucines have not been found in any fast loop while they show a high frequency in slow loops.

The three amino-acid types (Trp, Thr, His) representative for slow loops are polar, and they can participate in processes that involve the formation and breaking of hydrogen bonds occurring on the tens to hundreds nanosecond range. Bulky hydrophobic amino acids with aromatic side-chain rings more often appear in static loops than in mobile loops. Through the formation of disulfide bonds, cysteines tend to stabilize loop conformations rendering them static as is evident for 8 disulfide bonds, which belong to static loops in 5 proteins, out of a total number of 14 disulfide bonds in 7 proteins.

It is instructive to compare our loop time scale propensities with amino-acid "unfoldability" scales developed from databases of intrinsically disordered proteins (IDPs), such as DisProt[39,40] and TOP-IDP scales[41] (Figure S3). Although different unfoldability scales display significant differences between them, they show certain loose similarities to the loop time scale propensity scale of Figure S3. As a rule, amino acids that tend to be part of static (dynamic) loops have a lower tendency for unfolding (folding). Specifically, static amino acids, such as tyrosine, isoleucine, phenylalanine, and leucine, have a lower propensity for disorder and slow residues, such as threonine, valine, and histidine, all rank in the medium range of the disorder scales. However, the high unfoldability of lysine and aspartic acids does not match their propensity to belong to fast time scale loops. Also, significant discrepancies exist for methionine, cysteine, and tryptophan: while methionine ranks at the top of the list of fast loop residues, it has a relatively low probability for unfolding. Tryptophan, on the other hand, has the lowest propensity of all amino acids to belong to disordered

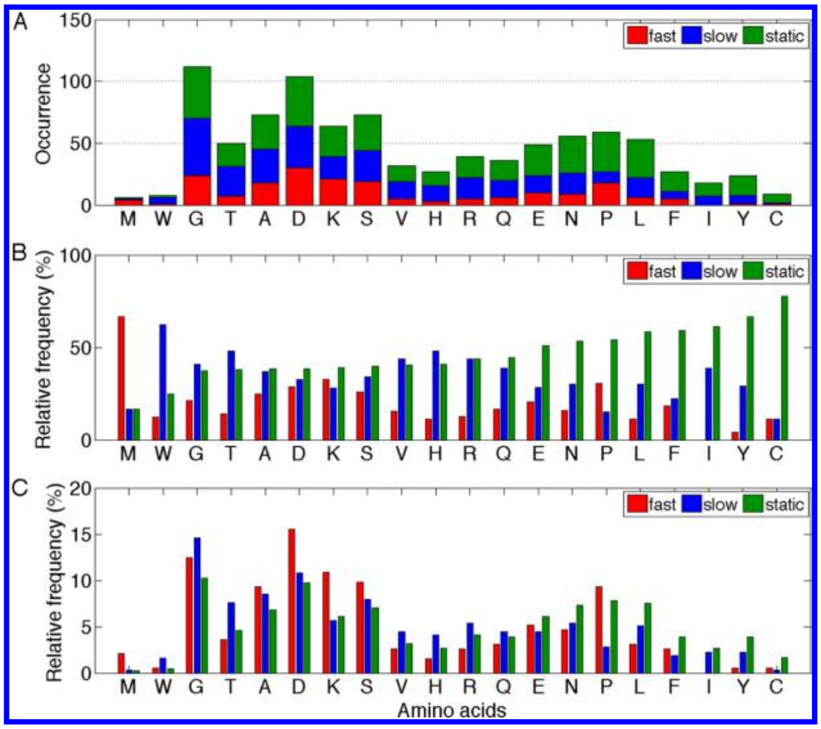


Figure 3. Absolute and relative frequencies of the 20 amino-acid types for 3 categories of loop motions: fast (red), slow (blue), static (green). (A) Distribution of all 1011 amino acids of the 38 proteins counting 169 loops for the 3 loop categories. (B) Probability distribution of each amino-acid type for the 3 loop categories. For each amino-acid type, the relative frequencies add up to 100%. (C) Distribution of 20 amino-acid types over the 3 loop categories. For each loop category the relative frequencies add up to 100%. Amino-acid names are indicated as 1-letter abbreviations, and the order is sorted based on the static loop frequency in panel B.

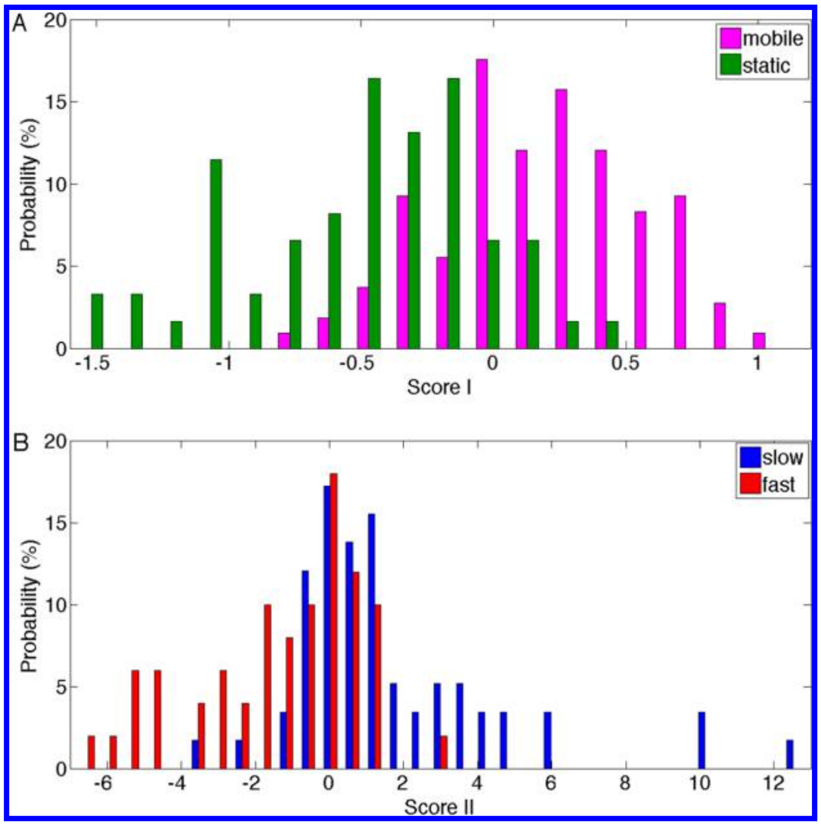


Figure 4. Performance of ToeLoop predictor for the two decision stages in the flowchart of Figure 1. (A) Distribution (histogram) of predicted scores to discriminate between static (green) and mobile (magenta) loops. The mobile loops comprise both slow and fast loops. The highest accuracy (81%) is achieved for the threshold score I of −0.33. (B) Distribution (histogram) of predicted score II to discriminate between slow (blue) and fast (red) loop categories. The highest accuracy (74%) is obtained for the threshold score of −0.98.

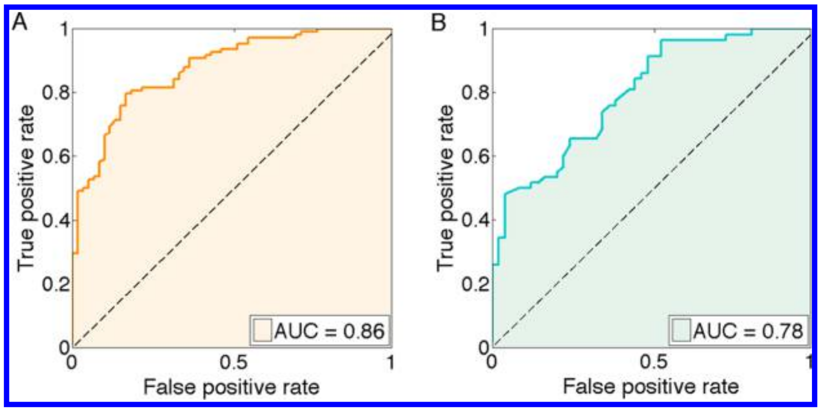


Figure 5. Receiver Operating Characteristic (ROC) performance of ToeLoop predictor. (A) ROC curve of ToeLoop to discriminate between mobile and static loops. The Area Under Curve (AUC) is 0.86. (B) ROC curve of ToeLoop to discriminate between slow and fast loops, which has an AUC of 0.78.

proteins, but when it is part of a loop, it has a high propensity to undergo slow dynamics.

To further improve the discrimination between the three types of protein loops, other chemical physical-chemical parameters complementing the primary sequence information have been incorporated. Histograms of six additional loop discriminators are shown in Figure S4. These parameters correspond to secondary structure affiliations (bridge, 3-helix, turn, or bend), loop lengths, loop contact sums, net electric charges of loops, average solvent accessibilities of loops, and average hydrophobicity indices of loops. The largest variations of these parameters are observed between fast and static loops. Fast loops are on average shorter, have more negatively charged residues, larger solvent accessibilities, and lower hydrophobicity indices and fewer contacts. Static loops, on the other hand, are on average longer, have lower solvent accessibilities, and higher hydrophobicity indices, and more contacts. Generally, these discriminators are somewhat more ambiguous to identify slow loops whose characteristics tend to fall between the fast and static categories.

3.3. Two-Stage Prediction of Loop Dynamics by ToeLoop. The ToeLoop predictor uses 13 amino-acid sequence based and biophysical descriptors reflecting the amino-acid composition, the secondary structure of loops, and tertiary contact sum between the loops and their environment to discriminate between mobile (fast or slow) and static loops. This distinction is performed by ToeLoop in the first stage (Figure 1) with the performance visualized in Figure 4A for all 169 protein loops. The Pearson correlation coefficient of predicted dynamics scores I vs true scores I is 0.64 (with all loops) and 0.57 (with leave-one-out cross-validation). The predicted mobility scores range from −1.5 to 1 with an overlap of the two major loop categories mostly in the −0.3 to 0.2 range. The prediction accuracy can be assessed by shifting the threshold score between the minimum and maximum score whereby a threshold value at −0.33 yields the highest accuracy of 81%. The receiver operating characteristic (ROC) curve of the first stage in ToeLoop is shown in Figure 5A. The large area under the curve (AUC) of 0.86 attests to the good performance of ToeLoop. To ensure an unbiased selection of static loops in the training set, a direct comparison of the histograms of predicted scores I from the two static loop sets is shown in Figure S5. The validation loop set yields an average prediction score I of −0.45 with 0.41 standard deviation, which closely matches the training set that has an average score of −0.50 with a standard deviation of 0.46.

Next, 108 loops classified as mobile in the first stage are categorized according to their time scales. Fast loops show significant averaging effects both for the Cα−Cβ order parameter and the chemical shifts on the sub-10 ns time scale, whereas slow loops display a high level of mobility with correlation times between 10 and 500 ns. This second classification stage is driven by amino-acid sequence information (see the Methods section). The predicted time scale score II is shown in the histogram of Figure 4B where the highest accuracy of 74% is reached for a threshold score II of −0.98. Moreover, the AUC belonging to the ROC curve of Figure 5B is 0.78, which indicates a good performance that is only slightly lower than the AUC of the first stage.

4. CONCLUSION

Since protein loops constitute the protein parts with the largest dynamic range, a predictive understanding of their mobility and associated time scales is important to elucidate their functional role. Analysis of protein loops in terms of their plasticity and time scales can provide important clues about their functional roles and the sequence of binding events during protein−protein recognition processes. In particular, fast loops have a relatively flat energy surface as compared to the thermal energy unit (k_BT), which can be easily reshaped by the presence of a binding partner and which is consistent with an induced fit mechanism.[42] Slow loops have a higher free energy barrier between different substates, which may not be sufficiently reshaped and lowered by the mere presence of the binding partner. This makes binding by conformational selection and population shift the preferred binding mechanism for such systems,[2,43] consistent with a model for protein binding that takes the binding time scales explicitly into account.[44]

We used MD simulations, validated by NMR chemical shift data, to probe multiple time scale dynamics of 169 protein loops belonging to 38 proteins. The improvement of chemical shift RMSDs when averaged over the full trajectories reflects good consistency with experiment. Furthermore, our statistical analysis reveals a direct relationship between loop dynamics time scales and amino-acid composition together with other biophysical loop descriptors, such as secondary structures, tertiary contacts, loop lengths, and solvent accessibilities. This relationship, which is coded in the Web server ToeLoop, allows the rapid prediction of loop time scales from an average protein structure. For the proteins used here we find a 81% prediction accuracy of static vs mobile loops and a 74% accuracy of fast vs slow loops. It permits the rapid prediction of loop behavior, which can be used to screen loops for specific properties in the context of functional studies as well as for *de novo* protein design.

■ ASSOCIATED CONTENT

Supporting Information

Additional tables and figures. Table S1: list of proteins used in this study and BMRB chemical shift assignment entries. Figure S1: effects of loop motions and time scales on NMR S^2 order parameters. Figure S2: change of average Cα and Cβ chemical shift RMSD upon increasing MD trajectory length. Figure S3: comparison of relative frequencies of 20 amino acids in the three loop categories with the unfoldability scales. Figure S4: distributions of six loop descriptors for the three loop categories. Figure S5: histogram comparison of ToeLoop predicted score I from two sets of static loops. Figure S6: Cα–Cβ order parameter dependence on loop category. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 614-688-2083. E-mail: bruschweiler.1@osu.edu. Corresponding author address: Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Huan-Xiang Zhou for discussions. This work was supported by the National Science Foundation (grant MCB-1360966) and the Ohio Supercomputer Center.

■ REFERENCES

(1) Fiser, A.; Do, R. K. G.; Sali, A. *Protein Sci.* **2000**, *9*, 1753–1773.
(2) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796.
(3) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 788–788.
(4) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.
(5) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.
(6) Cilia, E.; Pancsa, R.; Tompa, P.; Lenaerts, T.; Vranken, W. F. *Nucleic Acids Res.* **2014**, *42*, W264–W270.
(7) Cilia, E.; Pancsa, R.; Tompa, P.; Lenaerts, T.; Vranken, W. F. *Nat. Commun.* **2013**, *4*, 2741.
(8) Walsh, I.; Martin, A. J.; Di Domenico, T.; Tosatto, S. C. *Bioinformatics* **2012**, *28*, 503–509.
(9) Ishida, T.; Kinoshita, K. *Nucleic Acids Res.* **2007**, *35*, W460–W464.
(10) Yang, Z. R.; Thomson, R.; McNeil, P.; Esnouf, R. M. *Bioinformatics* **2005**, *21*, 3369–3376.
(11) Prilusky, J.; Felder, C. E.; Zeev-Ben-Mordehai, T.; Rydberg, E. H.; Man, O.; Beckmann, J. S.; Silman, I.; Sussman, J. L. *Bioinformatics* **2005**, *21*, 3435–3438.
(12) Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A. K. *Proteins* **2005**, *61* (Suppl7), 176–182.
(13) Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. *Bioinformatics* **2005**, *21*, 3433–3434.
(14) Halle, B. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 1274–1279.
(15) Zhang, F. L.; Brüschweiler, R. *J. Am. Chem. Soc.* **2002**, *124*, 12654–12655.
(16) Li, D. W.; Brüschweiler, R. *Biophys. J.* **2009**, *96*, 3074–3081.
(17) de Brevern, A. G.; Bornot, A.; Craveur, P.; Etchebest, C.; Gelly, J. C. *Nucleic Acids Res.* **2012**, *40*, W317–W322.
(18) Berjanskii, M. V.; Wishart, D. S. *J. Am. Chem. Soc.* **2005**, *127*, 14970–14971.
(19) Li, D. W.; Brüschweiler, R. *J. Chem. Theory Comput.* **2011**, *7*, 1773–1782.
(20) Berendsen, H. J. C.; van der Spoel, D.; Vandrunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.
(21) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
(22) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
(23) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
(24) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
(25) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
(26) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
(27) Prompers, J. J.; Brüschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 7305–7313.
(28) Prompers, J. J.; Brüschweiler, R. *J. Am. Chem. Soc.* **2002**, *124*, 4522–4534.
(29) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
(30) Li, D. W.; Brüschweiler, R. *J. Biomol. NMR* **2012**, *54*, 257–265.
(31) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321–333.
(32) Vila, J. A.; Ripoll, D. R.; Baldoni, H. A.; Scheraga, H. A. *J. Biomol. NMR* **2002**, *24*, 245–262.
(33) Vila, J. A.; Scheraga, H. A. *Acc. Chem. Res.* **2009**, *42*, 1545–1553.
(34) Li, D. W.; Brüschweiler, R. *J. Phys. Chem. Lett.* **2010**, *1*, 246–248.
(35) Markwick, P. R. L.; Cervantes, C. F.; Abel, B. L.; Komives, E. A.; Blackledge, M.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 1220–1221.
(36) Robustelli, P.; Stafford, K. A.; Palmer, A. G. *J. Am. Chem. Soc.* **2012**, *134*, 6365–6374.
(37) Gellman, S. H. *Biochemistry* **1991**, *30*, 6633–6636.
(38) Weininger, U.; Liu, Z.; McIntyre, D. D.; Vogel, H. J.; Akke, M. *J. Am. Chem. Soc.* **2012**, *134*, 18562–18565.
(39) Vucetic, S.; Obradovic, Z.; Vacic, V.; Radivojac, P.; Peng, K.; Iakoucheva, L. M.; Cortese, M. S.; Lawson, J. D.; Brown, C. J.; Sikes, J. G.; Newton, C. D.; Dunker, A. K. *Bioinformatics* **2005**, *21*, 137–140.
(40) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K. *Nucleic Acids Res.* **2007**, *35*, D786–D793.
(41) Campen, A.; Williams, R. M.; Brown, C. J.; Meng, J. W.; Uversky, V. N.; Dunker, A. K. *Protein Pept. Lett.* **2008**, *15*, 956–963.
(42) Koshland, D. E. *Angew. Chem., Int. Ed.* **1994**, *33*, 2375–2378.
(43) Csermely, P.; Palotai, R.; Nussinov, R. *Trends Biochem. Sci.* **2010**, *35*, 539–546.
(44) Zhou, H. X. *Biophys. J.* **2010**, *98*, L15–L17.

DOI: 10.1021/ct501085y