PCASSO: A Fast and Efficient Cα-Based Method for Accurately Assigning Protein Secondary Structure Elements

Sean M. Law,[a,b] Aaron T. Frank,[a,b] and Charles L. Brooks III[a,b]

Proteins are often characterized in terms of their primary, secondary, tertiary, and quaternary structure. Algorithms such as define secondary structure of proteins (DSSP) can automatically assign protein secondary structure based on the backbone hydrogen-bonding pattern. However, the assignment of secondary structure elements (SSEs) becomes a challenge when only the Cα coordinates are available. In this work, we present protein C-alpha secondary structure output (PCASSO), a fast and accurate program for assigning protein SSEs using only the Cα positions. PCASSO achieves ~95% accuracy with respect to DSSP and takes ~0.1 s using a single processor to analyze a 1000 residue system with multiple chains. Our approach was compared with current state-of-the-art Cα-based methods and was found to outperform all of them in both speed and accuracy. A practical application is also presented and discussed. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23683

Introduction

The basic protein secondary structure elements (SSEs), namely, α-helices and β-sheets, were first described by Pauling and Corey in 1951[1,2] and have since provided a foundation for comparing, classifying, and visualizing three-dimensional (3D) protein folds. Traditionally, protein SSEs were manually designated through visual inspection of the polypeptide chain, which often resulted in assignments that were subjective and, at times, incomplete. Today, this tedious process is made more efficient and reproducible through automated tools such as structural identification (STRIDE)[3] and define secondary structure of proteins (DSSP).[4,5] DSSP, one of the oldest and most popular SSE assignment programs available, assigns SSEs by first identifying all backbone carbonyl (C═O) and amide (N─H) hydrogen bonds based on a purely electrostatic criterion. Then, depending on the hydrogen bonding patterns, each residue is classified as a helix, strand, or loop. However, the assignment of SSEs becomes problematic when insufficient information is available [e.g., protein data bank (PDB) structures with unresolved backbone atoms, Cα-only models originating from cryo-electron microscopy (cryo-EM), and coarse-grained protein models used in multiscale simulations]. Although the positions of the missing backbone atoms that are required for SSE assignment can be estimated from reduced models,[6–11] the reconstruction methodology is imperfect and often requires some level of refinement or energy minimization through molecular dynamics (MDs) simulations to optimize the backbone hydrogen bonding networks before being processed through DSSP. Furthermore, this time consuming process can become prohibitive when reconstructing a large number of structures from long coarse-grained MD simulations. Thus, it is advantageous to develop a fast and efficient method that avoids the reconstruction process altogether and yet can still provide reliable SSE assignments that can be generally and consistently applied across multiple scales.

Several Cα-based assignment methods such as protein secondary element assignment (P-SEA),[12] voronoi tessellation assignment procedure (VoTAP), [13] and more recently, secondary structure assignment program based on only alpha carbons (SABA) [14] have been reported. P-SEA utilizes a combination of distances, angles, and dihedrals for secondary structure analysis while VoTAP generates contact matrices derived from 3D Voronoï tessellation, which are then used for assigning SSEs. SABA uses a similar approach to P-SEA but instead of directly computing the Cα coordinates SABA shifts the coordinates of the *i*th Cα atom to its pseudocenter (PC) position [defined as the center-of-geometry between Cα (*i*) and Cα ($i+1$)] and then assigns SSEs based on an optimized set of PC-dependent geometric criteria. This is thought to better represent the location of the backbone N─H/C═O atoms involved in secondary structure formation. While these methods appear to agree reasonably well with DSSP, P-SEA, and VoTAP are no longer being maintained and SABA is available only as a web server that is limited to analyzing individually uploaded PDB files.

In this work, we present protein C-alpha secondary structure output (PCASSO), a fast and efficient program for assigning protein SSEs that only requires Cα atoms as input. Using the well-known random forest (RF)[15] approach, PCASSO achieves

[a] *S. M. Law, A. T. Frank, Charles L. Brooks III*
Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109
[b] *S. M. Law, A. T. Frank, Charles L. Brooks III*
Department of Biophysics, University of Michigan, Ann Arbor, Michigan 48109, Fax: +1 (734) 647 1604, E-mail: brookscl@umich.edu
Contract/grant sponsor: NIH; contract/grant number: GM037554
Contract/grant sponsor: NSF (Center for Biological Physics); contract/grant number: PHY0216576
Contract/grant sponsor: University of Michigan President's Postdoctoral Fellowship (ATF)

high accuracy compared to DSSP and offers fast processing times even for large systems. PCASSO can be used for, but not limited to, evaluating individual PDBs, batch processing, and analyzing MD simulation trajectories. The source code (licensed under the GNU General Public License v3.0) and web server are made freely available at http://brooks.chem.lsa.umich. edu/software.

Methods

RF is an ensemble machine learning methodology that achieves high accuracy by aggregating classifications from independent random decision trees and reporting the mode vote.[15] To ensure that the trees within the forest are uncorrelated, each tree is trained on a bootstrap sample of the original data set (with replacement) and only a small, randomly chosen subset of features/variables is used to determine the best split at a given node. To compare our results with previous methods, we utilized the same protein training and test sets published by Mornon and coworkers[13] (see Supporting Information Tables S1–S3). All structural coordinates were obtained from the PDB[16] and analyzed with DSSP.[4,5] The Cα atoms were then extracted from each PDB and 258 basic geometric features (see below) were computed for each residue of the reduced model.

For a given residue, *i*, a set of features, $f_{C\alpha}(i)$ and $f_{PC}(i)$, were calculated from the Cα coordinates and the PC coordinates, respectively (see Supporting Information Table S4). The *j*th and *k*th residues form nonbonded interactions with the *i*th residue and help to identify interactions between strands that are separated in sequence. The *j*th residue has the shortest distance from residue *i* and, when *i* and *j* are from the same chain/segment, *j* must be at least $i + 6$ residues away. Similarly, the *k*th residue has the shortest distance from residue *i* and, when *i* and *k* are from the same chain/segment, *k* must be at least $i - 6$ residues away. The coordinates of the *i*th PC was previously defined as the center-of-geometry between Cα (*i*) and Cα $(i + 1)$[14] and so the PC coordinates for the last residue of each chain/segment is undefined as are the features that reference the ultimate C-terminal residue. The feature vector, *V*(*i*), for the *i*th residue is made up by features from the *i*th, $i - 1$th, and $i + 1$th residues (i.e. $V(i)=\{f_{C\alpha}(i), f_{PC}(i), f_{C\alpha}(i-1), f_{PC}(i-1), f_{C\alpha}(i+1), f_{PC}(i+1)\}$) which results in a total of $(2\times43)\times3=258$ feature elements.

From the training set, a total of 50 trees were generated using the RF implementation found in the Open Source Computer Vision (OpenCV) library[17] and default parameters were used unless otherwise specified. At each node, 16 out of 258 features/variables were selected at random to find the best split. Node splitting was ceased either when: (i) all members of the node were of the same class (i.e., helix, strand, or loop); (ii) the maximum depth allowed (25) was reached; or (iii) the minimum sample count required for a split (10) was not satisfied. Changes in the RF parameters (i.e., number of random features used for each split, maximum tree depth, minimum sample count, total number of trees, etc) did not result in a significant increase in accuracy. As the tree growing procedure is completely independent of the classification process, the resulting ensemble of trees was extracted from the OpenCV output, serialized as a string in preorder, and hardcoded into PCASSO for speed and efficiency. Thus, PCASSO is a standalone program that takes either PDB structures or MD simulation trajectories as input, deserializes the tree ensemble into independent binary decision trees, calculates the full feature vector for each Cα atom and processes it through each tree, aggregates the SSE classifications, and returns the mode vote for each residue of each structure or simulation snapshot. To compare the speed and accuracy of PCASSO with the reconstruction scheme, the missing backbone atoms for each Cα model from the test set were rebuilt using the *rebuild* program from the Multiscale Modeling Tools for Structural Biology Tool Set [6] and subsequently analyzed using DSSP. Finally, the protein test set was analyzed using PCASSO and the accuracy (relative to DSSP) was compared with the SSE assignments from P-SEA, VoTAP, and DSSP (using the reduced models with reconstructed backbone atoms as input). To demonstrate the value and applicability of PCASSO, we analyzed a previously published 58 μs MD folding trajectory of a human Pin1 WW domain variant called FiP35.[18] Simulation snapshots ($n = 2{,}900$) were assessed every 20 ns and the SSE classifications were used in constructing conformation space networks. All molecular graphics were generated in PyMOL[19] and SSE time series plots were created using in-house tools.

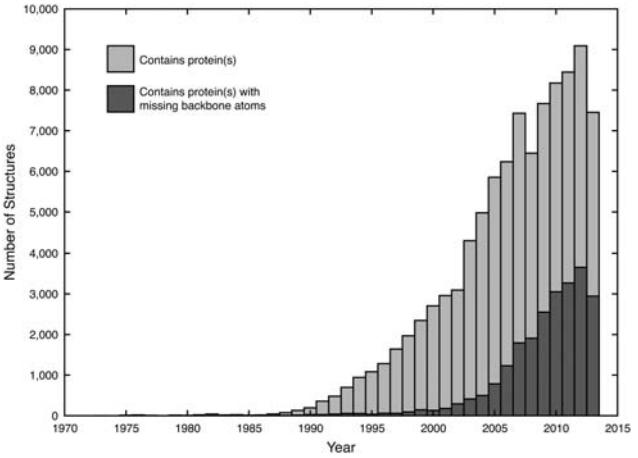

Figure 1. The number of protein-containing structures deposited in the PDB between 1971 and 2013 (noncumulative). A total of 96,286 PDB structures were analyzed and 23,295 PDB structures (~24%) were found to have incomplete/missing backbone atoms.

Results and Discussion

As the number of protein structures being deposited into the PDB grows, the number of X-ray, NMR, and cryo-EM structures with missing or incomplete backbone atoms also experiences a concomitant increase. For example, approximately 40% of the protein structures deposited in 2013 contained at least one or more missing backbone atoms (Fig. 1). Concurrently, the number of publications that include the terms "coarse," "grained," "protein," and "simulation" has also been on the rise.[20] As DSSP [4,5], the current gold standard for assigning SSEs, depends solely upon backbone hydrogen bonding patterns, residues with only Cα coordinates are generally ignored

Table 1. PCASSO accuracy comparison.

	Percent accuracy[a]			
SSE	PCASSO[b]	P-SEA[c]	VoTAP[d]	Reconstruction[e]
Helix	96.5 (96.6)	83.9	93.0	94.8
Strand	92.2 (95.3)	78.2	77.3	91.8
Loop	94.1 (92.2)	74.8	79.3	96.1
All	94.5	78.9	83.2	94.6

[a] DSSP is used as the reference. The true positive rate (sensitivity) is shown and the positive prediction value (precision) is in parentheses. [b] Trained on DSSP SSE assignments. [c] Computed using P-SEA (Ref. [12]). [d] Adapted from Ref. [13]. [e] See Methods

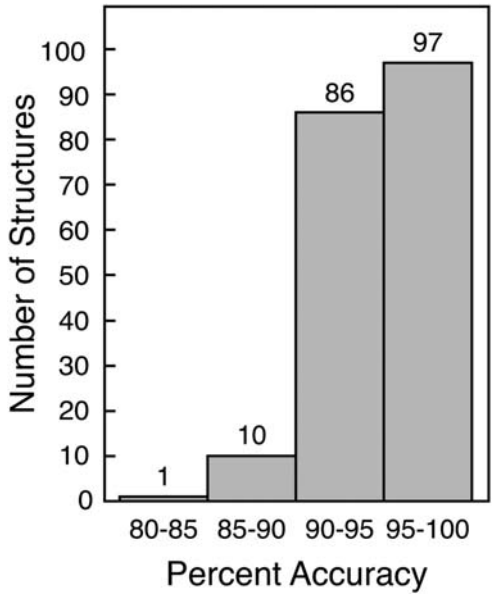

Figure 2. Histogram of structures with varying SSE assignment accuracies.

or neglected. While the backbone atoms for a single protein can be reconstructed from the Cα atoms with reasonable accuracy, this time-consuming process, as we will demonstrate below, becomes infeasible for much larger systems and/or for rapidly rebuilding a large ensemble of structures from coarse-grained/multiscale simulations. As scientists continue to push the size of systems that can be experimentally determined [21,22] or computationally simulated,[23] the demand for faster and more efficient analysis tools that can complement these larger systems will also rise. Thus, PCASSO has been developed to provide quick and reliable SSE classifications directly from the Cα coordinates (i.e., without backbone reconstruction) with the analogous aim of being to Cα-containing structures what DSSP is to all-atom structures.

To judge the performance of PCASSO, we compared our SSE assignment accuracy relative to DSSP with assignments from P-SEA and VoTAP (Table 1). Overall, PCASSO demonstrated ~95% accuracy, which is more than an 11% increase over P-SEA and VoTAP. PCASSO showed a substantial improvement in classifying strands and loops and a moderate enhancement in classifying helices. More importantly, PCASSO was found to be equally as accurate as the reconstruction scheme (i.e., the backbone atoms were reconstructed from the Cα coordinates and then evaluated using DSSP) and exhibited a high level of precision and sensitivity for each SSE class (i.e., low false positives and low false negatives). Over 94% of the structures in the test set had a greater than 90% classification accuracy and over 99% of the structures had a greater than 85% accuracy (Fig. 2). The three lowest accuracy structures (Supporting Information Table S5) only showed minor differences in their assignments and are displayed in Figure 3. Furthermore, as PCASSO was trained on DSSP SSE assignments, we also assessed the accuracy of PCASSO relative to STRIDE (Supporting Information Table S6). Remarkably, even without recalibrating PCASSO to match STRIDE, the overall accuracy was only slightly reduced to ~93% which can be attributed to a small decrease in accuracy for classifying helices and strands. It is logical that the accuracy results can somewhat vary when PCASSO is compared to different reference methods as STRIDE and DSSP are based on different approaches. In fact, it has been previously reported that STRIDE is in ~95% agreement with DSSP.[13] Additionally, it has been demonstrated that these minor discrepancies can be attenuated by the use of a ternary consensus method (TCM).[12,13,24] However, considering the generally high level of agreement with the aforementioned all-atom-based assignment methods, we contend that TCM would not be practical or necessary.

To assess the scalability of PCASSO, we evaluated its processing time for systems of increasing size using a single CPU (Table 2). We found that PCASSO was at least 24 times faster than P-SEA and at least 11 times faster than the reconstruction scheme. In fact, by extrapolation, as the number of residues (and/or structures) increases, it becomes infeasible to use any of the pre-existing Cα-based methods for assigning SSEs due to their much longer processing times. While in all cases, multiple structures or simulation snapshots can be divided amongst multiple CPUs in an "embarrassingly parallel" manner to boost the speed performance, only PCASSO is amenable to

Table 2. Comparison of SSE processing times.

			Time (s)						
PDBID	Residues	Chains	PCASSO	P-SEA	VoTAP	Reconstruction	DSSP	P-SEA PCASSO	Reconstruction + DSSP PCASSO
1PUC	101	1	0.01	0.34	–	0.11	0.04	34.00	15.00
1NBA	1011	4	0.11	2.74	–	1.04	0.17	24.91	11.00
1RVV	4620	30	1.25	51.39	–	12.75	1.17	41.11	11.14

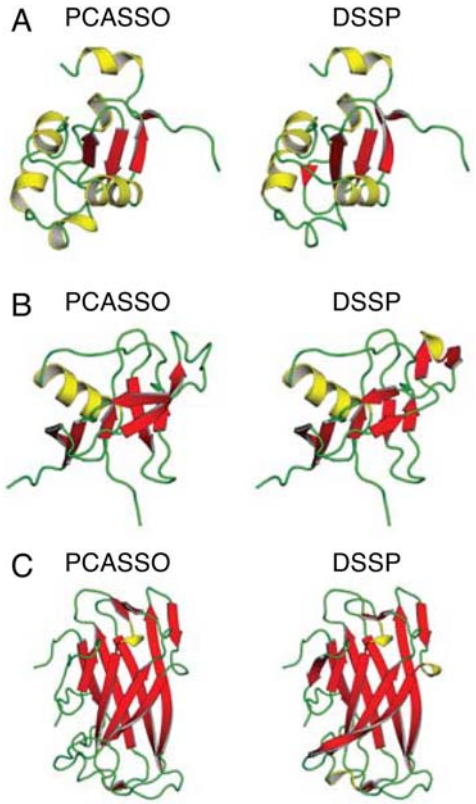

Figure 3. SSE assignment comparison for the three lowest accuracy structures.

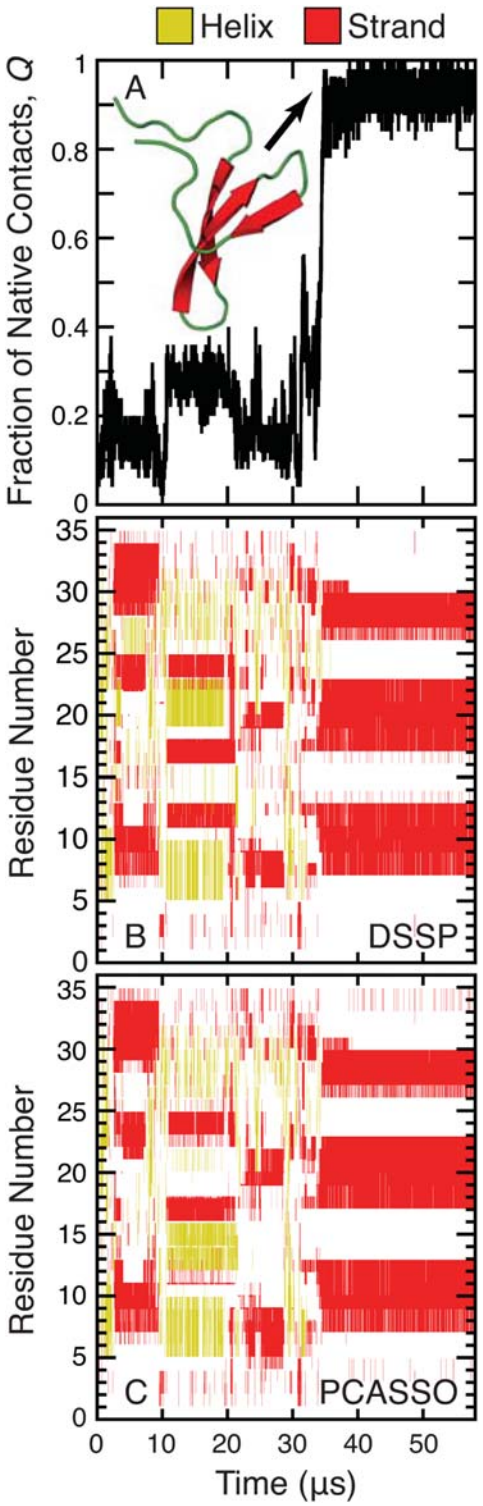

Figure 4. Analyses of the FiP35 folding trajectory.

further parallelization. For example, unlike P-SEA and VoTAP, which both assign helices first followed by strands and then loops (i.e., there is a residue assignment order dependency), PCASSO treats the assignment of each residue completely independently, which makes it perfectly suited for parallel processing. Additional speed improvements can also be made by distributing the evaluation of each independent decision tree to a different CPU or by removing redundant and/or highly correlated features. Thus, PCASSO is not only able to accomplish more with limited resources but its underlying implementation also allows room for future improvement and scalability.

The number of coarse-grained protein simulations has experienced a steady increase over the past decade as scientists seek to understand protein structure and dynamics on much longer timescales.[20] In the case of protein folding, the fraction of native amino acid contacts, Q,[25] is typically used as a progress variable for monitoring the folding process. However, Q can fail to identify important nonnative contacts or protein misfolding that would have otherwise been captured through SSE analysis. To illustrate this point and to demonstrate a practical application of PCASSO, we analyzed a previously published all-atom MD folding trajectory of a human Pin1 WW domain variant called FiP35,[18] which consists of a three-stranded β-sheet connected by two β hairpins (Fig. 4). Using Q as the reaction coordinate, initially, FiP35 is only partially folded but after $\sim$35 μs the peptide forms over 80% of its native contacts and is considered fully folded (Fig. 4A). However, both DSSP and PCASSO, which yield essentially the same results, reveal that FiP35 can form stable nonnative interactions at the onset and parts of the peptide actually misfold to a helix (Figs. 4B and 4C). Thus, this example clearly demonstrates the value of SSE assignments and how this information can be complementary to Q. Furthermore, PCASSO offers a

fast and reliable alternative to DSSP for analyzing protein secondary structure that can be applied to any Cα-containing multiscale model.

In conclusion, PCASSO outperformed pre-existing programs in both accuracy and speed. Given this, PCASSO can also be used in network analysis through SSE clustering,[26] high-throughput SSE studies, universal SSE assignments, SSE-based alignments,[27] renormalization of Gō-like models for intrinsically disordered proteins,[28] and to analyze coarse-grained simulation models that do not incorporate any native contact information[29,30] or where the native contacts are not known *a priori* (e.g., to examine cooperative folding of multimers or large multisubunit complexes). Ultimately, we hope that the work presented here will motivate the development of better and faster tools to complement the ever-growing challenges of big data.

Acknowledgments

The authors would like to acknowledge valuable scientific discussions with Bin Zhang, Junjie Zou, Shanshan Cheng, Shuai Wei, Logan Ahlstrom, Alex Dickson, Afra Panahi, Garrett Goh, and Karunesh Arora. The authors also thank D. E. Shaw Research for providing access to the FiP35 MD Trajectory.

Keywords: C-alpha models · secondary structure assignment · PCASSO · DSSP · STRIDE

How to cite this article: S. M. Law, A. T. Frank, C. L. Brooks *J. Comput. Chem.* **2014**, *35*, 1757–1761. DOI: 10.1002/jcc.23683

Additional Supporting Information may be found in the online version of this article.

[1] L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. USA* **1951**, *37*, 729.
[2] L. Pauling, R. B. Corey, H. R. Branson, *Proc. Natl. Acad. Sci. USA* **1951**, *37*, 205.
[3] D. Frishman, P. Argos, *Proteins: Struct. Funct. Genet.* **1995**, *23*, 566.
[4] R. P. Joosten, T. A. H. T. Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, G. Vriend, *Nucleic Acids Res.* **2011**, *39*, D411.
[5] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.
[6] M. Feig, J. Karanicolas, C. L. Brooks, III, *J. Mol. Graphics Modell.* **2004**, *22*, 377.
[7] Y. Q. Li, Y. Zhang, *Proteins* **2009**, *76*, 665.
[8] L. Holm, C. Sander, *J. Mol. Biol.* **1991**, *218*, 183.
[9] D. Petrey, Z. X. Xiang, C. L. Tang, L. Xie, M. Gimpelev, T. Mitros, C. S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I. Y. Y. Koh, E. Alexov, B. Honig, *Proteins: Struct. Funct. Genet.* **2003**, *53*, 430.
[10] P. Rotkiewicz, J. Skolnick, *J. Comput. Chem.* **2008**, *29*, 1460.
[11] A. Sali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779.
[12] G. Labesse, N. Colloch, J. Pothier, J. P. Mornon, *Comput. Appl. Biosci.* **1997**, *13*, 291.
[13] F. Dupuis, J. F. Sadoc, *J. P. Mornon Proteins* **2004**, *55*, 519.
[14] S. Y. Park, M. J. Yoo, J. Shin, K. H. Cho, *BMB Rep.* **2011**, *44*, 118.
[15] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.
[16] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[17] G. Bradski. *Dr Dobb's Journal of Software Tools*. **2000**, *25*, 120.
[18] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M .P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, W. Wriggers, *Science* **2010**, *330*, 341.
[19] Schrödinger (Schrödinger, LLC), *The Pymol Molecular Graphics System*.
[20] S. Takada, *Curr. Opin. Struct. Biol.* **2012**, *22*, 130.
[21] N. Volkmann, *Curr. Opin. Cell Biol.* **2012**, *24*, 141.
[22] Fridman K, Mader A, Zwerger M, Elia N, Medalia O, *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 736.
[23] M. Feig, Y. Sugita, *J. Mol. Graphics Model.* **2013**, *45*, 144.
[24] N. Colloch, C. Etchebest, E. Thoreau, B. Henrissat, J. P. Mornon, *Protein Eng.* **1993**, *6*, 377.
[25] E. Shakhnovich, G. Farztdinov, A. M. Gutin, M. Karplus, *Phys. Rev. Lett.* **1991**, *67*, 1665.
[26] F. Rao, A. Caflisch, *J. Mol. Biol.* **2004**, *342*, 299.
[27] Fontana P, E. Bindewald, S. Toppo, R. Velasco, G. Valle, S. C. Tosatto, *Bioinformatics* **2005**, *21*, 393.
[28] D. Ganguly, J. H. Chen, *Proteins: Struct. Funct. Bioinf.* **2011**, *79*, 1251.
[29] S. M. Gopal, S. Mukherjee, Y. M. Cheng, M. Feig, *Proteins* **2010**, *78*, 1266.
[30] P. Kar, S. M. Gopal, Y.-M. Cheng, A. Predeus, M. Feig, *J. Chem. Theory Comput.* **2013**, *9*, 3769.

Received: 22 April 2014
Revised: 5 June 2014
Accepted: 24 June 2014
Published online on 4 July 2014