Conformation Analysis

DOI: 10.1002/anie.201101641

Using Side-Chain Aromatic Proton Chemical Shifts for a Quantitative Analysis of Protein Structures**

*Aleksandr B. Sahakyan, Wim F. Vranken, Andrea Cavalli, and Michele Vendruscolo**

Chemical shifts are receiving renewed attention in structural biology owing to the recent introduction of novel methodologies that enable their use in protein structure determination.[1–10] As these approaches have so far been mostly concerned with backbone atoms, it would be highly desirable to further generalize them to also include side-chain atoms.[11–14] A major motivation for this objective is that side chains play crucial roles in determining the conformational properties of protein surfaces and interior cavities, which in most cases define the specificity of biomolecular interactions. In particular, aromatic side chains are capable of forming interactions with a variety of chemical groups through hydrophobic, π–π stacking, π–anion and π–cation interactions, and often comprise the hot spots of protein–protein[15] and protein–ligand[16] complex formation, and protein folding.[17] Furthermore, aromatic side chains, as sources of ring current effects, substantially influence the chemical shifts of other nuclei, including the highly exploited backbone nuclei. However, although ring-current terms are frequently included in chemical shift predictions of backbone nuclei, aromatic chemical shifts are not normally used to define the geometry of the aromatic rings themselves. Recent advances in specific labeling technologies for aromatic side chains[18,19] will soon increase the number of assigned aromatic chemical shifts, thus adding new prospects to the established methodology of aromatic chemical shift measurements.[20] The incorporation of chemical shifts of aromatic side chains in structure-determination algorithms, in addition to the backbone atoms, would make it possible to extend the use of chemical shifts in structural studies. To achieve this goal, a chemical shift prediction method for side-chain nuclei that is based solely on the configurations of proximal atoms needs to be developed.[21]

This type of predictions, which is at variance with other currently available chemical shift predictors that provide chemical shift evaluations for side-chain nuclei,[22–25] are readily differentiable with respect to the atomic coordinates, and thus enable the calculation of biasing forces for the integration of the equations of motion within a molecular dynamics scheme.[8] Prediction of aromatic side-chain chemical shifts by differentiable functions opens new opportunities to monitor a range of important processes, and will increase the scope of chemical shift usage in determining the structures of biomolecular complexes and complex biomolecular systems.[3,6]

To address this challenge, we present here ArShift, a chemical shift prediction method for protein side-chain aromatic ^{1}H nuclei. We then demonstrate that by using only aromatic side-chain chemical shifts, structures that do not match the state from which chemical shifts are measured can be revealed. The ArShift predictions are based on known phenomenological terms that describe the effects of ring current,[26] magnetic anisotropy,[27] and electric field[28] terms, which are accompanied by a set of dihedral angle terms and distance-based polynomials[21] (see the Supporting Information). A comprehensive analysis of the aromatic chemical shift assignments available from the BMRB database[29] is used after filtering and re-referencing steps[30] to reduce the number of inaccurate and artifactual entries (Figures S1 and S2 in the Supporting Information). To identify the mapping between chemical shifts and structures, only structures determined by X-ray crystallography at a resolution of 2.0 Å or better are considered in the derivation of the geometric terms. The combination of terms used in the predictions is then optimized through a Monte Carlo approach to decrease the number of fitted coefficients, thus increasing the significance of the remaining ones (Table S1).

We assessed the accuracy of the prediction method by performing individual predictions (in leave-one-out tests) for all the chemical shift entries used for deriving the coefficients. The standard deviations of the residual errors (denoted here as standard errors) for the models implemented in the ArShift package are 0.189, 0.204, 0.256, 0.191, and 0.173 ppm for Phe-^{1}Hδ, Phe-^{1}Hε, Phe-^{1}Hζ, Tyr-^{1}Hδ, and Tyr-^{1}Hε nuclei, respectively (Figures S3 and S4). The comparison of the ArShift standard errors and the standard deviations of the corresponding chemical shift types in the BMRB database are presented in Figure 1.

Predictions for ^{13}C nuclei are not reported in this work because they do not currently provide a significant improvement over those based on the average values derived from the BMRB database. The reason for this situation is most probably the neglect of the stronger isotope effects on ^{13}C

[*] A. B. Sahakyan, Dr. A. Cavalli, Prof. M. Vendruscolo
Department of Chemistry, University of Cambridge
Lensfield Road, Cambridge CB2 1EW (UK)
E-mail: mv245@cam.ac.uk
Dr. W. F. Vranken[+]
European Bioinformatics Institute
Wellcome Trust Genome Campus
Cambridge CB10 1SD (UK)
[+] Current Address: Structural Biology Brussels
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussel (Belgium)

[**] This research was supported by the Herchel Smith Foundation, the Leverhulme Trust, EMBO, the BBSRC, the Royal Society, and the EU eNMR project no. 213010.
Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201101641.

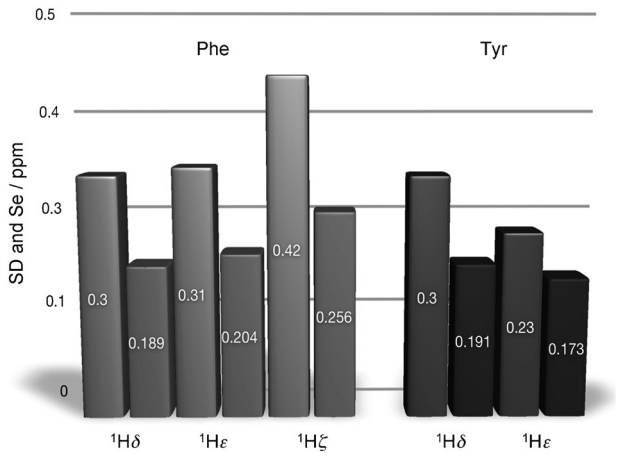

Figure 1. Performance of the ^{1}H chemical shift predictions for different types of protein aromatic side-chain nuclei. For each atom type the bars on the right show the standard errors (ppm) of the ArShift method. The bars on the left show the standard deviations of the corresponding chemical shifts in the BMRB database.

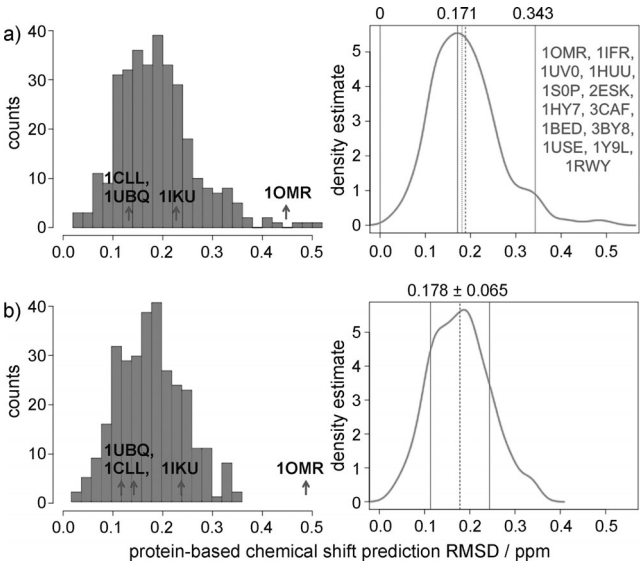

Figure 2. Accuracy of the ArShift predictions in terms of RMSD distributions (ppm) from the protein-based leave-one-out tests. Results before (a) and after (b) exclusion of 13 outlier structures from the total of 452 structures in the parameterization.

nuclei caused by the immediately attached nuclei. It will perhaps become possible to account for these effects in the parameterization step by considering a database that, besides the chemical shift values, includes information about the isotopic state of the attached hydrogen atoms (that is, deuterated or not).

We then performed a protein-based leave-one-out test, in which we removed individual protein entries from the model development data set in turn, and then predicted all the corresponding chemical shifts. The protein-based root-mean-square deviation (RMSD) of ArShift calculated in this way is (0.178 ± 0.065) ppm. In order to increase the accuracy of the predictions, we used a self-consistent approach in which the ArShift model optimization and parameterization were carried out twice. After the initial model generation, the examination of the RMSD distribution from the protein-based leave-one-out test (Figure 2a) revealed the existence of a high-RMSD shoulder next to the normal distribution of RMSD values centered at around 0.171 ppm. Thus, all the structures that fall outside two standard deviations were further examined, and we found that in all these cases the structures derived from X-ray crystallography were substantially different from those derived from NMR spectroscopy because of significant conformational changes upon Ca^{2+} ion or ligand binding, or sequence alterations (Figure 3). Some X-ray structures were also lacking peptide segments that were present in the corresponding NMR structures (light-blue moieties in Figure 3). Therefore, even though all the structures used in the parameterization process were determined in the crystal form, the first iteration of the model generation process resulted in a predictor that self-diagnosed the cases where the crystal structures did not match those in solution for which chemical shifts had been measured. This finding demonstrates that the high-resolution structures derived from X-ray crystallography used for development of the predictor do train coefficients that are not biased towards crystal structures.

After removal of the 13 proteins for which the ArShift predictions detected mismatches between structures derived from X-ray crystallography and NMR spectroscopy, a second iteration of model optimization and parameterization was carried out with the remaining 439 high-resolution structures determined by X-ray crystallography, in order to generate the final predictor.

To further illustrate the applicability of the ArShift predictor, we analyzed the 2K39[32] and 2NR2[33] ensembles, and the 1D3Z[34] set of structures in comparison to the 1UBQ[35] structure of ubiquitin derived from X-ray crystallography (Figure S5–S7). The results indicate that the 1D3Z structure is the most consistent with the experimental aromatic side-chain ^{1}H chemical shifts, followed by 1UBQ, 2NR2, and 2 K39 (Figures 4, S6, and S7). This result illustrates the importance of using RDC side-chain measurements as restraints to determine the positions of the side chains with high accuracy, as was the case for the 1D3Z structures.[34]

A similar test for a calmodulin structure derived from X-ray crystallography (1CLL[31]) and a solution-state ensemble (1X02[18]) highlights the overall good quality of the former structure as an average representation of the structure of this protein (Figures 5, S8, and S9), because the aromatic side-chain ^{1}H chemical shifts back-calculated from the 1CLL structure are in very good agreement with the experimental chemical shifts.[18]

We also found that averaging the predicted aromatic chemical shifts over the 20 conformers in the 1X02 solution-state ensemble improves the agreement between the predicted and experimental chemical shift values. An obvious exception from this trend is Phe-89, thus suggesting the presence of a possible imprecision in the structure or in the dynamics of this particular residue in the 1X02 ensemble.

A comparison with other existing prediction methods[22–25] illustrates the excellent performance of ArShift (Figures S10–

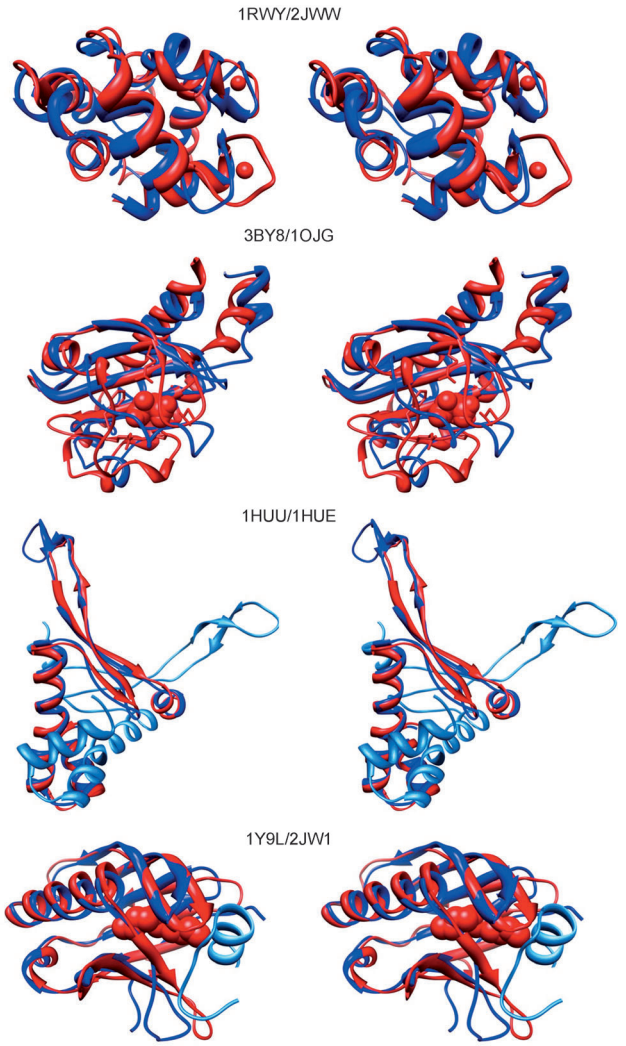

Figure 3. Stereoview of representative cases identified by ArShift, in which structures derived from X-ray crystallography (red) and NMR spectroscopy (blue) differ significantly, for example because of Ca^{2+} ion or ligand binding, or missing segments.

12). A test on recoverin[36,37] in its Ca^{2+}-bound and free states, which substantially differ in their conformations, indicate that ArShift is more sensitive towards structural imperfections than the other methods that we considered (Figure S11).

The ArShift predictor can be used in structure calculations to score conformations on the basis of their consistency with measured chemical shifts. We illustrate this aspect in the cases of the 124-residue DNA binding domain of SV40 T-antigen and the 56-residue protein GB3. The DNA-binding domain of SV40 T-antigen contains 10 Phe and 7 Tyr residues, of which 37 aromatic ^{1}H chemical shifts are available;[38] these chemical shifts also meet the filtering criteria described above. The 2FUF X-ray structure,[39] for which use of ArShift results in predictions with 0.161 ppm RMSD (Figure S13), was used as a starting point for a 17 ns molecular dynamics trajectory to sample the conformational space of this protein domain (see Figure S14 and Methods section in the Supporting Information). From the resulting trajectory, we analyzed 2430 structures (extracted at 7 ps intervals) by using ArShift.

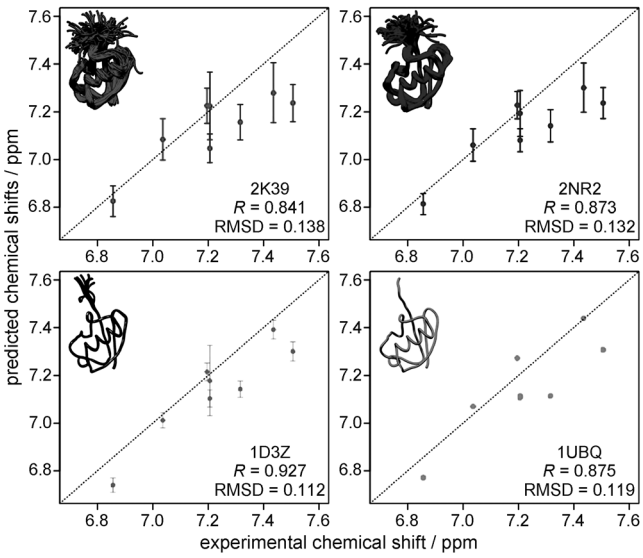

Figure 4. Correlation between predicted and experimental ^{1}H chemical shifts for the Phe and Tyr side chains in three solution-state ensembles (2K39,[32] 2NR2,[33] and 1D3Z[34]) and structure of ubiquitin (1UBQ[35]) derived from X-ray crystallography. Standard deviations of the predicted chemical shift values over multiple conformers are shown as error bars. Pearson correlation coefficients (R) and RMSDs (ppm) are also shown. Codes refer to PDB entries.

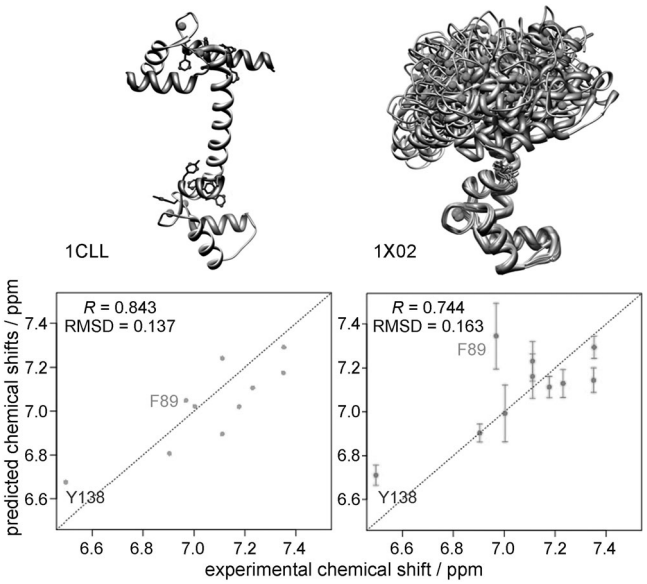

Figure 5. Correlation between predicted and experimental aromatic ^{1}Hε chemical shifts for an X-ray crystal structure (1CLL[31]) and a solution-state ensemble (1X02[18]) of calmodulin. Standard deviations of the corresponding predicted chemical shift values over the conformers in the solution-state ensemble are shown as error bars. Pearson correlation coefficients (R) and RMSDs (ppm) are also shown.

In the cases of both the DNA-binding domain of SV40 T-antigen and GB3, the scoring function defining the agreement between predicted and experimental chemical shifts with respect to the structural RMSD from its native state

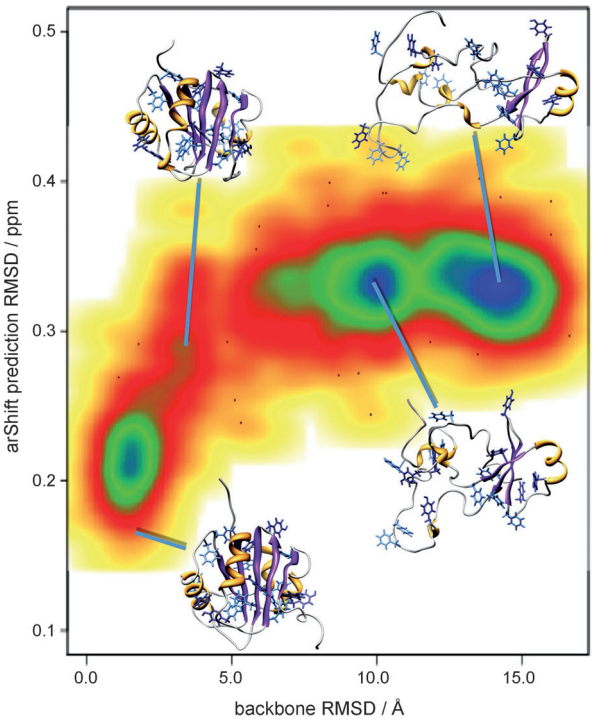

Figure 6. ArShift prediction RMSDs plotted against the backbone structural RMSDs; 2430 structures of the DNA-binding domain of SV40 T antigen were used to draw this plot. The color code indicates the density of the data points; 4 representative structures and 25 points from the lowest-density areas are explicitly shown.

(Figures 6, S15, and S16) is highly funneled, as required in structure calculations.

We anticipate that the ArShift method will be constantly improved as more experimental chemical shift measurements will become available in the BMRB and other databases. The ArShift method is available both as a stand-alone code and as an application on the web (http://www-vendruscolo.ch.cam.ac.uk/software.html and Figure S17).

Received: March 7, 2011
Published online: September 2, 2011

Keywords: chemical shift prediction · conformation analysis · NMR spectroscopy · peptides · protein structures

[1] A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 9615 – 9620.
[2] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. H. Liu, A. Eletsky, Y. B. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, A. Bax, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4685 – 4690.
[3] R. W. Montalvao, A. Cavalli, X. Salvatella, T. L. Blundell, M. Vendruscolo, *J. Am. Chem. Soc.* **2008**, *130*, 15990 – 15996.
[4] P. Robustelli, A. Cavalli, M. Vendruscolo, *Structure* **2008**, *16*, 1764 – 1769.
[5] M. Berjanskii, P. Tang, J. Liang, J. A. Cruz, J. J. Zhou, Y. Zhou, E. Bassett, C. MacDonell, P. Lu, G. H. Lin, D. S. Wishart, *Nucleic Acids Res.* **2009**, *37*, W670 – 677.
[6] R. Das, I. Andre, Y. Shen, Y. B. Wu, A. Lemak, S. Bansal, C. H. Arrowsmith, T. Szyperski, D. Baker, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18978 – 18983.
[7] F. A. A. Mulder, M. Filatov, *Chem. Soc. Rev.* **2010**, *39*, 578 – 590.
[8] P. Robustelli, K. Kohlhoff, A. Cavalli, M. Vendruscolo, *Structure* **2010**, *18*, 923 – 933.
[9] D. M. Korzhnev, T. L. Religa, W. Banachewicz, A. R. Fersht, L. E. Kay, *Science* **2010**, *329*, 1295 – 1296.
[10] D. S. Wishart, *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *58*, 62 – 87.
[11] R. E. London, B. D. Wingad, G. A. Mueller, *J. Am. Chem. Soc.* **2008**, *130*, 11097 – 11105.
[12] F. A. A. Mulder, *ChemBioChem* **2009**, *10*, 1477 – 1479.
[13] D. F. Hansen, P. Neudecker, L. E. Kay, *J. Am. Chem. Soc.* **2010**, *132*, 7589 – 7591.
[14] D. F. Hansen, P. Neudecker, P. Vallurpalli, F. A. A. Mulder, L. E. Kay, *J. Am. Chem. Soc.* **2010**, *132*, 42 – 43.
[15] P. B. Crowley, A. Golovin, *Proteins* **2005**, *59*, 231 – 239.
[16] C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061 – 5084.
[17] B. S. Frank, D. Vardar, D. A. Buckley, C. McKnight, *J. Protein Sci.* **2002**, *11*, 680 – 687.
[18] M. Kainosho, T. Torizawa, Y. Iwashita, T. Terauchi, A. M. Ono, P. Güntert, *Nature* **2006**, *440*, 52 – 57.
[19] P. Lundström, P. Vallutpalli, D. F. Hansen, L. E. Kay, *Nat. Protoc.* **2009**, *4*, 1641 – 1648.
[20] C. Redfield, F. M. Poulsen, C. M. Dobson, *Eur. J. Biochem.* **1982**, *128*, 527 – 531.
[21] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, M. Vendruscolo, *J. Am. Chem. Soc.* **2009**, *131*, 13894 – 13895.
[22] X. P. Xu, D. A. Case, *J. Biomol. NMR* **2001**, *21*, 321 – 333.
[23] J. Meiler, *J. Biomol. NMR* **2003**, *26*, 25 – 37.
[24] J. Lehtivarjo, T. Hassinen, S. P. Korhonen, M. Peräkylä, R. Laatikainen, *J. Biomol. NMR* **2009**, *45*, 413 – 426.
[25] B. Han, Y. Liu, S. W. Ginzinger, D. S. Wishart, *J. Biomol. NMR* **2011**, *50*, 43 – 57.
[26] C. W. Haigh, R. B. Mallion, *Prog. Nucl. Magn. Reson. Spectrosc.* **1979**, *13*, 303 – 344.
[27] K. Ösappy, D. A. Case, *J. Am. Chem. Soc.* **1991**, *113*, 9436 – 9444.
[28] A. D. Buckingham, *Can. J. Chem.* **1960**, *38*, 300 – 307.
[29] J. L. Markley, E. L. Ulrich, H. M. Berman, K. Henrick, H. Nakamura, H. Akutsu, *J. Biomol. NMR* **2008**, *40*, 153 – 155.
[30] W. Rieping, W. F. Vranken, *Proteins* **2010**, *78*, 2482 – 2489.
[31] R. Chattopadhyaya, W. E. Meador, A. R. Means, F. A. Quiocho, *J. Mol. Biol.* **1992**, *228*, 1177 – 1192.
[32] O. F. Lange, N. A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, B. L. de Groot, *Science* **2008**, *320*, 1471 – 1475.
[33] B. Richter, J. Gsponer, P. Varnai, X. Salvatella, M. Vendruscolo, *J. Biomol. NMR* **2007**, *37*, 117 – 135.
[34] G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, *J. Am. Chem. Soc.* **1998**, *120*, 6836 – 6837.
[35] S. Vijay-Kumar, C. E. Bugg, W. J. Cook, *J. Mol. Biol.* **1987**, *194*, 531 – 544.
[36] T. Tanaka, J. B. Ames, T. S. Harvey, L. Stryer, M. Ikura, *Nature* **1995**, *376*, 444 – 447.
[37] O. H. Weiergräber, I. I. Senin, P. P. Philippov, J. Granzin, K. W. Koch, *J. Biol. Chem.* **2003**, *278*, 22972 – 22979.
[38] X. Luo, D. G. Sanford, P. A. Bullock, W. W. Bachovchin, *Nat. Struct. Biol.* **1996**, *3*, 1034 – 1039.
[39] G. Meinke, P. A. Bullock, A. Bohm, *J. Virol.* **2006**, *80*, 4304 – 4312.