Conformational Dynamics

DOI: 10.1002/anie.201209669

Mapping the Population of Protein Conformational Energy Sub-States from NMR Dipolar Couplings**

*Paul Guerry, Loïc Salmon, Luca Mollica, Jose-Luis Ortega Roldan, Phineus Markwick, Nico A. J. van Nuland, J. Andrew McCammon, and Martin Blackledge**

The precision with which X-ray crystallography and nuclear magnetic resonance (NMR) have provided structural models of biologically active and inactive conformations of countless proteins belies an easily overlooked dilemma. Proteins are inherently dynamic, exhibiting conformational freedom on timescales from picoseconds to seconds, implicating structural rearrangements that are essential for their biological function.[1] Classical structural biology determines static models, that afford little insight into the underlying conformational equilibrium. The role that structural dynamics play in biological processes can only be understood by characterizing all thermally accessible protein conformations and their populations.[2]

NMR spectroscopy is uniquely sensitive to the presence of conformational dynamics in solution. Residual dipolar couplings (RDCs) measured in weakly aligned proteins,[3,4] scalar couplings, and chemical shifts,[5–7] probe motions occurring on timescales faster than 100 s of microseconds. These parameters therefore offer general tools to characterize protein motion on physiologically important timescales.[8–10]

A common approach to the dynamic interpretation of RDCs is to combine experimental restraint terms with a classical potential-energy force field to develop a conformational ensemble in agreement with experimental data.[11–14] RDCs have also been exploited to characterize the conformational space sampled by the protein backbone either by fitting experimental data to determine angular excursions of internuclear bond vectors,[15–18] or in comparison with different levels of accelerated molecular dynamics (AMD)[19,20] to describe the most appropriate ensemble.[21,22] Comparison of motions modeled using the Gaussian axial fluctuation (GAF) model, with ensembles derived from restraint-free AMD, demonstrated that such methods can provide a convergent description of protein motion.[18,23]

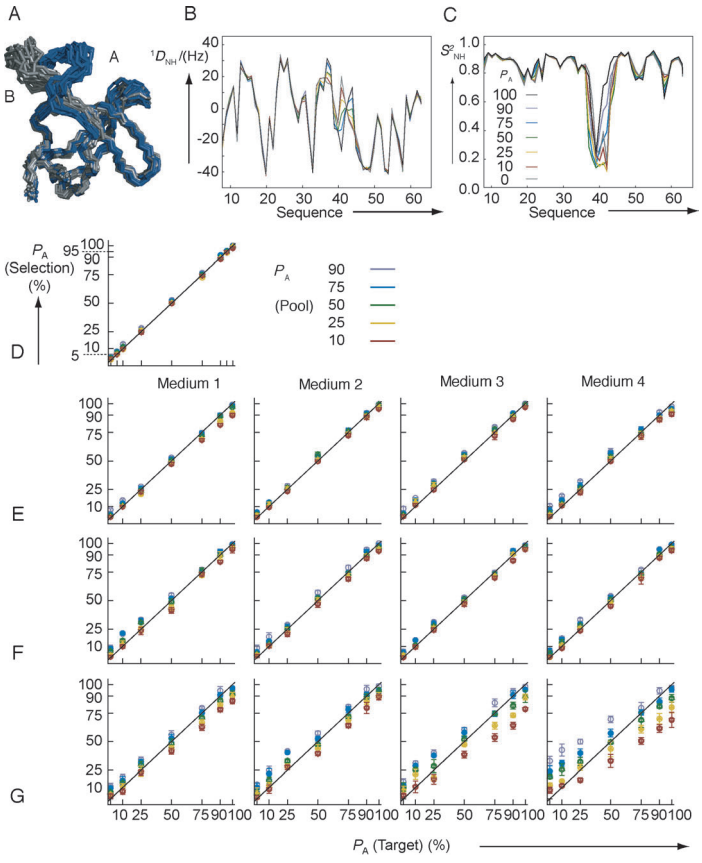

Figure 1. Determination of sub-state populations using SUPERNOVA. A) The model system—an SH3 domain—exhibits two distinct sub-states A and B from which B) noise-modulated RDCs were simulated and mixed in varying proportions. C) S^2_{NH} values for the different targets with P_A from 0–100%. D–G) Proportion of A in the selection, as a function of the proportion in the target. Colors indicate the proportion present in the pool. D) Using $^1D_{NH}$, $^1D_{CC}$, $^2D_{CHN}$, and $^1D_{CN}$ from four orthogonal alignment tensors. E) Using $^1D_{NH}$, $^1D_{CC}$, $^2D_{CHN}$, and $^1D_{CN}$ from one tensor (applied to each of four tensors). F) Only $^1D_{NH}$ from three of the tensors (four combinations are shown, the named medium is not used). G) Only $^1D_{NH}$ from one tensor (four tensors are shown).

[*] Dr. P. Guerry,[+] Dr. L. Salmon,[+] Dr. L. Mollica, Dr. M. Blackledge
Protein Dynamics and Flexibility
Institut de Biologie Structurale Jean-Pierre Ebel
CNRS-CEA-UJF UMR 5075, 41 rue Jules Horowitz
38027 Grenoble Cedex (France)
E-mail: martin.blackledge@ibs.fr

Dr. P. Markwick, Prof. J. A. McCammon
Department of Chemistry and Biochemistry UCSD San Diego CA
Howard Hughes Medical Institute
San Diego Supercomputer Center, La Jolla CA (USA)

Prof. N. A. J. van Nuland
Structural Biology Brussels, Vrije Universiteit Brussel
Pleinlaan (Belgium)

Dr. J.-L. Ortega Roldan
Department of Biochemistry, University of Oxford
South Parks Road, Oxford (UK)

[+] These authors contributed equally to this work.

[**] Financial support from the CEA, CNRS, UJF, ANR-12-BS07-0023-01 ComplexDynamics (M.B.), and the CCRT is acknowledged. The work at UCSD was supported by the NSF, NIH, CTBP, NBCR, and NSF Supercomputers.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201209669.

The general application of these approaches is hindered by the amount of experimental data that are required to accurately describe the motion. For many experimental systems it is not possible to measure RDCs in a sufficiently diverse combination of alignment media to allow the application of analytical approaches. Comparison with AMD, on the other hand, supposes that simulations carried out at the "optimal" level of acceleration correctly describe all accessible conformations and their populations. Existing protocols[18,21,22] are therefore poorly adapted to the quantification of modulations in the Boltzmann ensemble as a function of environmental conditions or the presence or absence of a partners.

We present a general method that exploits experimental RDCs to map the free-energy landscape occupied by folded proteins in solution, determining populations of accessible conformational sub-states contributing to the dynamic equilibrium. The method initially exploits multi-level AMD simulation, flooding the conformational space available to the protein as completely as possible to sample different sub-states, which are combined to provide an extensive pool of conformers, comprising both high-and low-energy conformations. Boltzmann-weighted ensembles are then assembled by comparison with experimental NMR data. Ensemble selection is achieved using model-free interpretation of RDCs combined with a specifically designed genetic algorithm. The approach is termed SUPERNOVA (sub-state populations on potential-energy surfaces using restraints from NMR spectroscopy and conformational oversampling).

The accuracy of SUPERNOVA, and its robustness against bias in the pools of structures from which the ensembles are selected, were tested using a synthetic dataset simulated from a hypothetical system. Two 200 ns MD simulations of the SH3C domain from CD2 AP were performed (see the Supporting Information). The first, referred to as "A", sampled the native fold, and the second, "B", an alternative conformation, that was adapted in the loop region from residues 32 to 44 (Figure 1). 1000 structures were retained from each trajectory (target A and target B) from which RDCs were simulated, while 10000 different structures from the same trajectories constituted sampling pools (pool A and pool B) for selection. Conformers from target A and target B were combined with populations $P_A = 100, 90, 75, 50, 25, 10, 0\,\%$ and $P_B = (100 - P_A)\,\%$. Noise-modulated RDCs were simulated from these targets using four orthogonal alignment tensors (supporting information). Variation of RDCs and order parameters (S^2_{NH}) between targets are shown in Figure 1.

SUPERNOVA selections aimed at the seven targets described above were initiated from seven different pools, containing the varying proportions of pools A and B. The results are summarized in Figure 1, showing the proportion of selected structures having the A or B loop conformation as a function of the pool composition. For calculations performed with a full set of RDCs ($^1D_{NH}$, $^1D_{CC}$, $^2D_{CHN}$ and $^1D_{CN}$ from all four media), the proportion of A-type structures selected matches the target to within $\pm 1\,\%$ for all targets and pools. Additional tests show similar accuracy for populations as low as 5 % (Figure 1 D). Figure 2 illustrates typical selected ensembles, in the $\{\phi/\psi\}$ space, using principal component (PC) analysis. SUPERNOVA correctly and reproducibly identifies

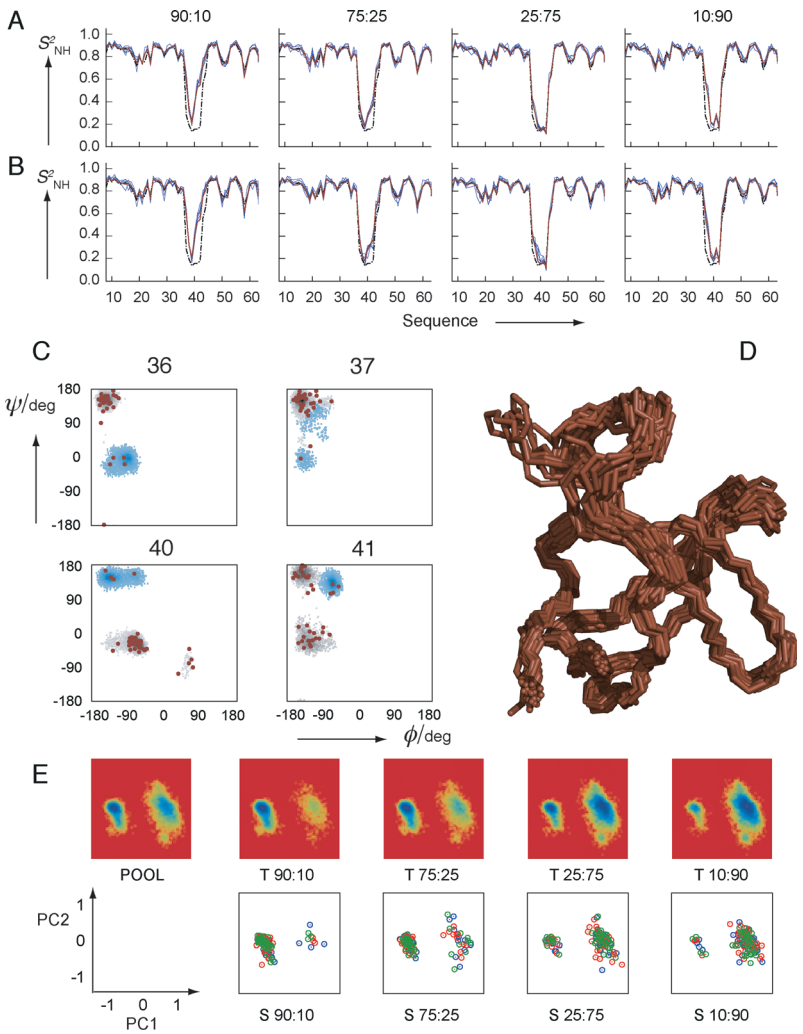

Figure 2. Reproduction of synthetic dynamics and populations. A) S^2_{NH} for different targets (proportions of A and B shown above the panels) using $^1D_{NH}$, $^1D_{CC}$, $^2D_{CHN}$, and $^1D_{CN}$ from four orthogonal alignment tensors (the dashed line represents the pool) S^2_{NH} from target (red), S^2_{NH} from five independent selections (blue). B) As (A) using only $^1D_{NH}$ from three tensors. C) ϕ/ψ distributions for ensemble selected against target with proportions 90:10 (red) compared to trajectory A (blue) and trajectory B (gray). The proportion of A and B was 10:90 in the pool. 4/40 structures are within the $\{\phi/\psi\}$ regions specific to conformation B. D) Ensemble selected against 90:10 target from 10:90 pool. E) PC analysis of conformational ensembles. Top line: 50:50 pool, and four targets (proportions as shown). Blue indicates highest values. Bottom line three distinct selections for the four different targets (in this case four couplings were used from four media).

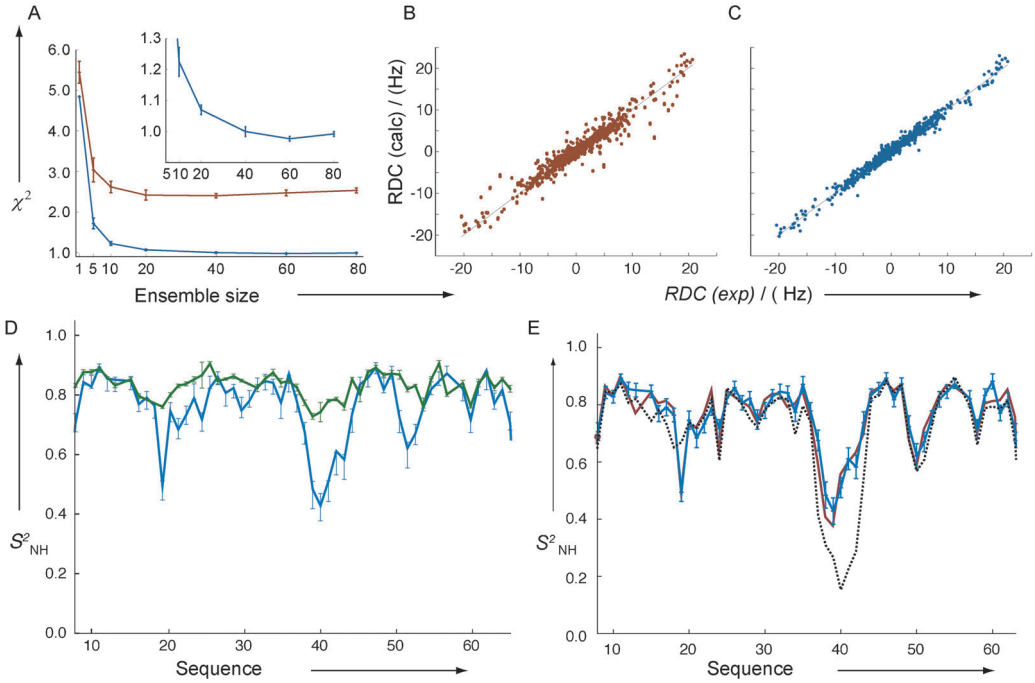

Figure 3. Application of SUPERNOVA to describe the conformational equilibrium of the SH3C from CD2AP. A) Estimation of the optimal number of structures in the selected ensemble. Blue: reduced χ^2 for RDCs measured in passive alignment media (mean and standard deviation over 50 independent calculations; inset: zoom of the same curve). Red: random selections of the same number of conformers from the pool. B,C) Comparison of back-calculated RDCs from the pool (red) and those predicted from SUPERNOVA ensembles (blue). D) Comparison of fast (picoseconds to nanoseconds) S^2_{NH} determined from relaxation (green) and S^2_{NH} from SUPERNOVA reporting on motions up to milliseconds (blue). Error bars from noise-based MC simulations. E) SUPERNOVA S^2_{NH} determined using all RDCs from all data sets (blue), and three sets of NH RDCs from three most independent alignment media (red), compared to S^2_{NH} calculated over the entire sampling pool (black, dotted line).

populations of major and minor species (90:10) in spite of unfavorable initial bias (10:90). Repetition using sparser data identifies populations as low as 10% using only $^1D_{NH}$ from three media, or four RDCs from one medium. $^1D_{NH}$ from one medium is insufficient to accurately determine sub-state populations (Figure 2E).

Application of SUPERNOVA to experimental data requires that the ensemble of conformers in the sampling pool adequately covers the conformational space available to the protein. To address this challenge, we initially identify the level of acceleration that best reproduces experimental data, using published AMD protocols.[18,22] Additional sampling is then achieved using higher levels of acceleration. Distinct conformational sub-states in the potential-energy landscape are identified from all AMD calculations, from which standard MD simulations are carried out, ensuring the energetic viability of all conformers in the pool, as well as allowing analysis of fast motional properties. Conformational sub-states that are populated in different proportions to those present at the optimum level of acceleration, are then combined as a function of the experimental data.

The selection protocol was applied to an extensive set of experimental RDCs measured from the protein SH3C from CD2AP dissolved in 10 alignment media.[18] The optimal number of structures within the ensemble (40 in this case) was determined using systematic cross-validation, removing one entire dataset from each of 10 analyses (Figure 3a). Figure 3 shows S^2_{NH} of selection and pool, compared to fast (<4 ns) motions determined from spin relaxation.[24] Critically, the analysis was repeated using only $^1D_{NH}$ from three alignment media, giving identical S^2_{NH} profiles (Figure 3E), confirming that the procedure is robust when fewer data are available. Statistical uncertainty in the ensemble description was assessed using noise-based Monte–Carlo (MC) simulations. This uncertainty is expressed in terms of error bars of S^2_{NH} or dispersion in PC or dihedral angle analysis (Figures 3 and 4).

The n-SRC loop, comprising residues 36 to 42, exhibits the most motion and shows clear modulation of the population of different sub-states relative to the pool ensemble for residues 37 and 40. Residues 18 and 19 also show population redistribution, resulting in a more flexible equilibrium than present in the pool (Figure 3). $\{\phi/\psi\}$ distributions of the three regions in which slow motions of significant amplitude are present again highlight population redistribution (Figure 4). PC analyses of pool and selections (the three highest populated modes account for 30% of the total motional amplitude) maps the accessible potential-energy landscape (Figure 4). Populations of different sub-states can be estimated on this basis (see Table S2 and Figure S5 in the Supporting Information). Representative structures of the major conformation, and two weakly populated sub-states are shown in Figure 4. PC analysis also identifies the most significant component of collective motion, involving the same loops (19–20 and 37–43).

Figure 4 shows a representative SUPERNOVA ensemble of 40 structures, indicating the position of the three regions exhibiting large-amplitude slow motions. The conformational sampling occurring within three of the most populated sub-states is also shown, corresponding to motions sampled on timescales up to the nanosecond range. Interconversion between the sub-states evidently occurs on the slower timescales (nanosecond to microsecond scales) defined by RDCs and chemical shifts. This separation of timescales is an essential feature of AMD-based RDC analysis, allowing for

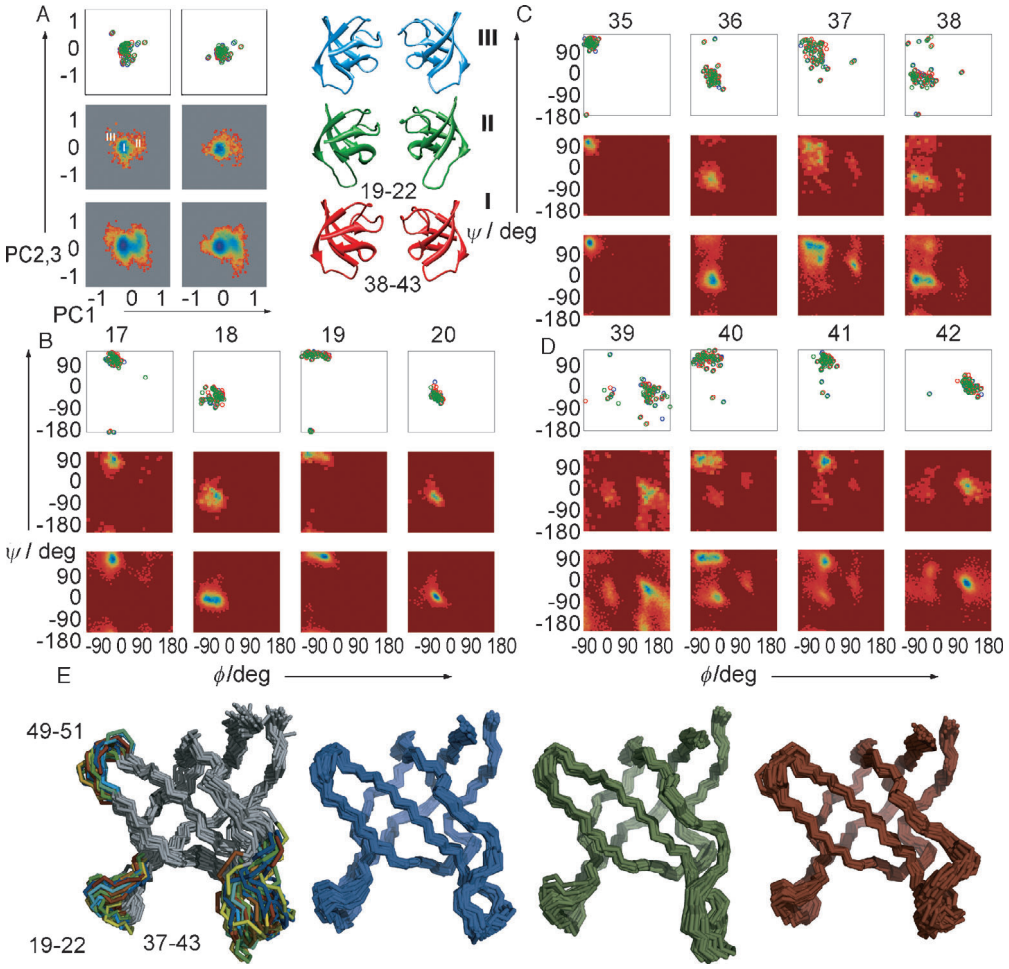

Figure 4. Mapping populations of conformational sub-states in CD2AP SH3C. A) PC analysis. Left and right figures show second and third most populated components (*y*-axis) against most significant component (*x*-axis). Bottom panels: PC analysis of the pool. Middle panels: PC analysis of 100 noise-based Monte Carlo simulations. Top panels: individual components of three independent ensembles (green, red, and blue). Representative structures from the three regions shown in the PCA plot of the SUPERNOVA selection are shown on the right (rotated 180° about the vertical axis). B–D) ϕ/ψ distributions from 12 amino acids. Bottom panels—distribution over the pool, middle panels, populations derived from 100 Monte Carlo simulations. Scale runs from dark red (negligible population) through yellow to dark blue (most populated). Top panels: individual ϕ/ψ values of three independent ensembles (green, red, and blue). E) Left: SUPERNOVA ensemble of 40 structures. Colored loops show regions with significant slow (nanoseconds to milliseconds) dynamics. Fast dynamic (picoseconds to nanoseconds) amplitudes occurring within three distinct sub-states selected from the ensemble (blue, green, and red).

the differentiation between motional averaging regimes that are specific to each NMR parameter.[18,21,22]

Reproduction of entire RDC datasets from different media not included in the selection improves significantly compared to those calculated from the pool (Figure 3). Similarly ^{15}N, $^{1}H^{N}$, $^{13}C^{\alpha}$, and $^{13}C^{\beta}$ backbone chemical shifts calculated using the program SHIFTX[25] and averaged over the ensembles are significantly better reproduced by the RDC-based selection than the pool. All chemical shifts from the loop regions exhibiting slower motions show improved reproduction compared to the pool and to the individual structure that best fits all RDCs (average improvements of 22 and 30%, respectively; see the Supporting Information). SUPERNOVA was also tested against an alternative pool composition, derived from three different fully solvated MD simulations (a total of 1.5 μs). This pool samples conformational space differently for a number of amino acids, but converges to a very similar selection, in terms of {ϕ/ψ} distribution, PC analysis, and S^2_{NH} (Supporting Information).

The growing volume of direct experimental evidence indicating the importance of weakly populated states for the function and malfunction of proteins,[26] highlights the importance of developing methods that can provide a quantitative molecular description of all sub-states present in solution. The SUPERNOVA approach addresses this central question by combining solutions to four principal difficulties that have hindered progress in the resolution of quantitative conformational sub-state populations.

Ensemble analysis of RDCs has suffered from the difficulty of predicting the physical alignment of the protein in all but the simplest cases of steric alignment.[27] We bypass this problem, generating a common mathematical solution reporting on all alignments, but more importantly, providing a measure of data reproduction by the ensemble of conformers that is essential to the selection algorithm.[28] Secondly, we present an approach that is demonstrably robust at manageable experimental costs, with as few as three sets of $^{1}D_{NH}$ or different peptide plane RDCs from a single medium, both achievable for a large number of proteins. Thirdly, a genetic algorithm combines members of a large ensemble of energetically viable conformers into sub-ensembles in agreement with the data,[29] avoiding the unpredictable effects of combining experimental and physical terms in a hybrid force field. Finally, the structural pool is generated using an enhanced sampling approach, based on multi-level AMD simulation, in which conformational space is over-sampled following an initial comparison to experimental data. Conformers from all simulations are combined into a pool from which sub-ensembles are selected, allowing for modulation of the

rugosity of the potential-energy landscape as a function of experimental data.

SUPERNOVA remains accurate in the presence of relatively sparse experimental data, allowing the application of cross-validation procedures that demonstrate its predictive power. Crucially, the limits of the resolution of each experimental dynamic description can be readily established a posteriori, using MC noise-based analyses as shown in the application to SH3C from CD2AP. In this case SUPERNOVA identifies the presence of motions in the nanosecond to millisecond range, in three surface loops, two of which are involved in recognition of physiological partners.[24] Future comparison of slow dynamic interconversion between substates in the free form of the protein with those present in physiological complexes will provide fascinating insight into their role in molecular recognition.

In conclusion, the combination of RDCs with innovative approaches to the efficient sampling of conformational space and specifically designed ensemble selection, provides access to dynamic averaging occurring on timescales extending from the picosecond to the millisecond. RDCs characterize the presence of dynamics on an amino-acid-specific basis, whereas AMD and ensemble matching map these motions within the accessible conformational space. This combination is predictive of independent experimental data, and provides a significantly better description of protein dynamics than static or alternative dynamic descriptions. The method is shown to provide statistically meaningful ensemble descriptions of protein motions, and is viable using experimental NMR measurements that are accessible for a large population of soluble proteins, ensuring a broad applicability of the approach to the study of physiologically relevant protein dynamics.

Received: December 3, 2012
Published online: February 1, 2013

Keywords: conformational sampling · dipolar couplings · molecular dynamics · NMR spectroscopy · proteins

[1] M. Karplus, J. Kuriyan, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6679 – 6685.
[2] H. Frauenfelder, S. Sligar, P. Wolynes, *Science* **1991**, *254*, 1598 – 1603.
[3] N. Tjandra, A. Bax, *Science* **1997**, *278*, 1111 – 1114.
[4] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, J. H. Prestegard, *Nat. Struct. Biol* **1997**, *4*, 292 – 297.
[5] M. V. Berjanskii, D. S. Wishart, *J. Am. Chem. Soc.* **2005**, *127*, 14970 – 14971.
[6] D.-W. Li, R. Brueschweiler, *J. Phys. Chem. Lett.* **2010**, *1*, 246 – 248.
[7] P. R. L. Markwick, C. F. Cervantes, B. L. Abel, E. A. Komives, M. Blackledge, J. A. McCammon, *J. Am. Chem. Soc.* **2010**, *132*, 1220 – 1221.
[8] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, R. Brüschweiler, *J. Am. Chem. Soc.* **2001**, *123*, 6098 – 6107.
[9] J. Tolman, K. Ruan, *Chem. Rev.* **2006**, *106*, 1720 – 1736.
[10] T. S. Ulmer, B. E. Ramirez, F. Delaglio, A. Bax, *J. Am. Chem. Soc.* **2003**, *125*, 9179 – 9191.
[11] G. M. Clore, C. D. Schwieters, *Biochemistry* **2004**, *43*, 10678 – 10691.
[12] K. Lindorff-Larsen, R. Best, M. DePristo, C. Dobson, M. Vendruscolo, *Nature* **2005**, *433*, 128 – 132.
[13] O. F. Lange, N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, B. L. de Groot, *Science* **2008**, *320*, 1471 – 1475.
[14] R. B. Fenwick, S. Esteban-Martin, B. Richter, D. Lee, K. F. A. Walter, D. Milovanovic, S. Becker, N. A. Lakomek, C. Griesinger, X. Salvatella, *J. Am. Chem. Soc.* **2011**, *133*, 10336 – 10339.
[15] G. Bouvignies, P. Bernadó, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, M. Blackledge, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13885 – 13890.
[16] G. Bouvignies, P. Markwick, R. Brüschweiler, M. Blackledge, *J. Am. Chem. Soc.* **2006**, *128*, 15100 – 15101.
[17] L. Salmon, G. Bouvignies, P. Markwick, N. Lakomek, S. Showalter, D.-W. Li, K. Walter, C. Griesinger, R. Brüschweiler, M. Blackledge, *Angew. Chem.* **2009**, *121*, 4218 – 4221; *Angew. Chem. Int. Ed.* **2009**, *48*, 4154 – 4157.
[18] L. Salmon, L. Pierce, A. Grimm, J.-L. O. Roldan, L. Mollica, M. R. Jensen, N. van Nuland, P. R. L. Markwick, J. A. McCammon, M. Blackledge, *Angew. Chem.* **2012**, *124*, 6207 – 6210; *Angew. Chem. Int. Ed.* **2012**, *51*, 6103 – 6106.
[19] A. F. Voter, *Phys. Rev. Lett.* **1997**, *78*, 3908 – 3911.
[20] P. R. L. Markwick, J. A. McCammon, *Phys. Chem. Chem. Phys.* **2011**, *13*, 20053 – 20065.
[21] P. R. L. Markwick, G. Bouvignies, M. Blackledge, *J. Am. Chem. Soc.* **2007**, *129*, 4724 – 4730.
[22] P. R. L. Markwick, G. Bouvignies, L. Salmon, J. A. McCammon, M. Nilges, M. Blackledge, *J. Am. Chem. Soc.* **2009**, *131*, 16968 – 16975.
[23] L. Salmon, G. Bouvignies, P. Markwick, M. Blackledge, *Biochemistry* **2011**, *50*, 2735 – 2747.
[24] L. Salmon, J.-L. Ortega Roldan, E. Lescop, A. Licinio, N. van Nuland, M. R. Jensen, M. Blackledge, *Angew. Chem.* **2011**, *123*, 3839 – 3843; *Angew. Chem. Int. Ed.* **2011**, *50*, 3755 – 3759.
[25] S. Neal, A. Nip, H. Zhang, D. S. Wishart, *J. Biomol. NMR* **2003**, *26*, 215 – 240.
[26] A. J. Baldwin, L. E. Kay, *Nat. Chem. Biol.* **2009**, *5*, 808 – 814.
[27] M. Zweckstetter, *Nat. Protoc.* **2008**, *3*, 679 – 690.
[28] S. A. Showalter, R. Brüschweiler, *J. Am. Chem. Soc.* **2007**, *129*, 4158 – 4159.
[29] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, M. Blackledge, *J. Am. Chem. Soc.* **2009**, *131*, 17908 – 17918.

