THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Exploring Regions of Conformational Space Occupied by Two-Domain Proteins

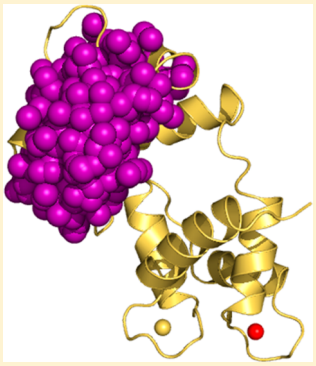
Witold Andrałojć,[†,‡] Claudio Luchinat,*[,†,‡] Giacomo Parigi,[†,‡] and Enrico Ravera[†,‡]

[†]Center for Magnetic Resonance, University of Florence, Via L. Sacconi 6, 50019, Sesto Fiorentino, Italy

[‡]Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3, 50019, Sesto Fiorentino, Italy

Supporting Information

ABSTRACT: The presence of heterogeneity in the interdomain arrangement of several biomolecules is required for their function. Here we present a method to obtain crucial clues to distinguish between different kinds of protein conformational distributions based on experimental NMR data. The method explores subregions of the conformational space and provides both upper and lower bounds of probability for the system to be in each subregion.



1. INTRODUCTION

Many biologically relevant processes are made possible by the existence of at least one of the partners in multiple conformational states.[1−10] Besides the biological relevance, these systems are also a benchmark for the development of experimental biophysical methods.[11−14] In this context, several approaches based on the data-driven creation of optimized conformational ensembles have been proposed and applied to a number of biological systems.[15−25]

Especially in the presence of large mobility, all such methods fall short in representing the "real" conformational heterogeneity,[26,27] given that the number of possible conformations is much larger than the number of experimental observables and, in any case, completely unknown. In mathematics, such situations go under the name of ill-conditioned and ill-posed inverse problems and, as such, present an infinite number of solutions. In more general terms, this means that several combinations of the possible conformations can reproduce the experimental data. This intrinsic limitation is compounded with the presence of the experimental error, which broadens even more the spectrum of the available solutions. In the presence of more restricted mobility, the situation is rather different, and the simple model-free methods are usually employed to grasp the information about the residual mobility.[28−30]

In this paper we propose a method to evaluate the size and shape of the conformational space sampled by a two-domain protein and to discover, without any assumption, whether it switches or not between structurally very different conformations. The method is based on the definition of regions of the conformational space and allows us to assess whether the experimental data can be completely accounted for, for instance, by excluding a given region or conversely whether conformations residing in such region must be necessarily included.[31]

Once a region is defined, it is possible to calculate both the upper and lower occurrence limits for that region (maxOR and minOR) as the largest and smallest weight, respectively, that all conformations belonging to such region must have to provide averaged data in agreement with the experimental data, once complemented by other conformations outside this region. The procedure is similar to a previously published procedure to assess the maximum weight of individual conformations.[32] Notably, the minimum occurrence of a region (minOR) can be immediately obtained from the difference between 1 and the maxOR of the complementary region. The possibilities offered by this approach are here analyzed and discussed through simulated examples where paramagnetic NMR data are used as synthetic restraints mimicking the experimental data; the approach is also used to analyze the conformational variability of calmodulin when bound to a peptide derived from the death-associated protein kinase 1 (DAPk1) protein using experimental paramagnetic NMR data.

2. METHODS

MaxOR and minOR calculations are performed by searching ensembles of protein conformations that comply with experimental data, by imposing that a subset of these conformations (i) belongs to a previously defined region of the conformational space and (ii) is sampled at the desired

Received: May 16, 2014
Revised: August 20, 2014
Published: August 21, 2014

weight. The maxOR and minOR values for this region are defined as the largest and smallest weight that can be given to such subset of conformations without worsening the agreement with the experimental data. The observables considered in the present study are paramagnetic-NMR based restraints: pseudocontact shifts (PCSs), and self-orientation residual dipolar couplings (RDCs).[12,13,32−34]

A large pool of sterically allowed conformations must be first generated and the observables associated with each conformation calculated. In the present case, M = 50 000 conformations of the protein calmodulin (CaM) were generated using the program RanCh,[15] and Np = 194 PCSs and Nr = 134 RDCs induced by the presence of three lanthanide ions (terbium(III), thulium(III), and ytterbium(III)) were calculated for each conformation. The residues for which the PCS and RDCs were calculated were selected on the basis of ref 34 (the data set relative to the complex with the DAPk1 peptide). The RDCs used are fewer than the PCS, given their remarkable dependence on the local mobility, which affects to a lesser extent the PCS.[35]

Optimized ensembles of protein conformations were searched through a minimization providing the best possible agreement between the weighted average of PCSs and RDCs calculated for the conformations of the ensemble and the experimental data. As already indicated, the ensemble was constrained to comprise a subset of conformations (taken from the pool) belonging to a defined region with a given total weight, and it was completed by other conformations (each of them with its own weight but with the constraint that the weight of all conformations in the ensemble equals 1) selected from the pool outside the defined region. Several calculations must be performed by changing the given weight of the conformations belonging to the defined region. The properties of the ensembles generated at each minimization step are irrelevant as long as structures belonging to the defined region with a given total weight are contained therein.

To allow for fast calculations, the minimization was implemented so as to be based on a regularized linear problem. The $M \times$ Np matrix containing the Np calculated PCSs for each of the M structures of the pool and the $M \times$ Nr matrix containing the Nr calculated RDCs for the same structures were divided by the norm of the experimental data and then combined into a matrix $A = M \times$ (Np + Nr). Also the experimental PCS and RDC data were normalized by their norm and combined into a vector b. The agreement between experimental and back-calculated data corresponds to the Q factor,[36] calculated as

$$Q = \| Aw - b \|_2 \tag{1}$$

A frugal coordinate descent algorithm, combined with random coordinate search,[37] was used to solve the regularized linear system,

$$\text{argmin}\{\| Aw - b \|_2^2 + \lambda(1 - \sum_{i=1}^{N} w_i)^2\}, \quad \text{s.t.w.} > 0 \tag{2}$$

where the sum of the weights of all conformations in the ensemble is constrained to 1. This minimization was performed to determine the lowest possible Q value.

The maxOR and minOR values were determined by solving

$$\text{argmin}\left\{\| Aw - b \|_2^2 + \lambda[(w_{MO} - \sum_{i=1}^{N\in C} w_i)^2 - (1 - w_{MO} - \sum_{i=1}^{N\in D} w_i)^2]\right\}, \quad \text{s.t.w.} > 0 \tag{3}$$

where w_{MO} is the fixed value that must correspond to the sum of the weights of all conformations within the predefined region and where C and D indicate the structures within and outside that region, respectively. Again, the largest w_{MO} providing a good fit of the experimental data (with Q below 1.2 of the lowest possible Q value) defines the maxOR of the region, whereas the smallest w_{MO} defines its minOR.

To define the regions in the conformational space to be analyzed, good starting points can be the structures with largest maximum occurrence (MaxOcc).[38] The latter is defined as the maximum occurrence, i.e., as the maximum weight, that a single conformation can have whatever ensemble it belongs to,[32] and it can be calculated with eq 3 where C comprises the selected structure and D all of the rest of the pool.

3. RESULTS AND DISCUSSION

To characterize the conformational variability of a protein with some degree of internal flexibility, we define here the concepts of maxOR and minOR as the maximum and minimum occurrence of regions defined in the conformational space of the protein. MaxOR and minOR are thus the maximum and minimum percent of time, respectively, that the protein can spend in any ensemble of conformations belonging to a defined region and still be in agreement with the experimental data. MaxOR and minOR can thus provide precious information for the characterization of the conformational variability of systems composed of two rigid domains connected by a flexible linker.

Before the analysis of an experimental data set, the method and its possibilities were assessed by three simulations. We have simulated the case of limited interdomain flexibility, the case of two-site exchange, and the case of large interdomain flexibility. Simulations 2 and 3 are compared in terms of recovery of the original distributions when the latter was generated with more or less the same extent of averaging of the experimental data. The protocols used in the calculations are detailed below. The first simulation (limited interdomain flexibility) is actually modeled to mimic the case of CaM bound to a peptide derived from death-associated protein kinase 1 (DAPk1),[34] which we further proceed to analyze. CaM is a ubiquitous and highly conserved calcium-binding protein composed of two domains connected by a flexible linker, which allows the protein to sample different conformations depending on the reciprocal domain positions. The PCSs and RDCs, induced by the presence of three lanthanide ions (terbium(III), thulium(III), and ytterbium(III)) selectively substituted to the calcium ion in the second calcium binding site of the N-terminal domain of the protein,[39] were used as restraints.

It is important to remark that different experimental data sets can be used to define MaxOcc and the related quantities, as proven by the inclusion of SAXS and PRE in the calculations presented in previous papers. This leads to two further considerations: (a) the first is how to add relevant information, i.e., to add different experimental observables that have different dependence on the conformational properties; (b) the second is how many experimental points are needed for each data set. Such analysis has been elegantly performed by Berlin et al.[16]

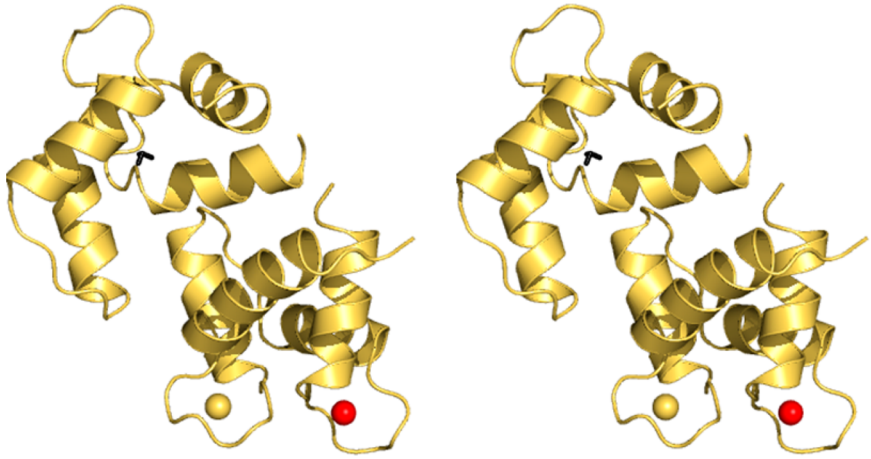


Figure 1. Stereoview (cross-eye) of the conformation selected for the generation of synthetic data. The red and yellow spheres represent the calcium-substituted lanthanide and the calcium ions, respectively. This conformation can be represented by a triad of Cartesian axes (in black), centered at the center of mass of the C-terminal domain.

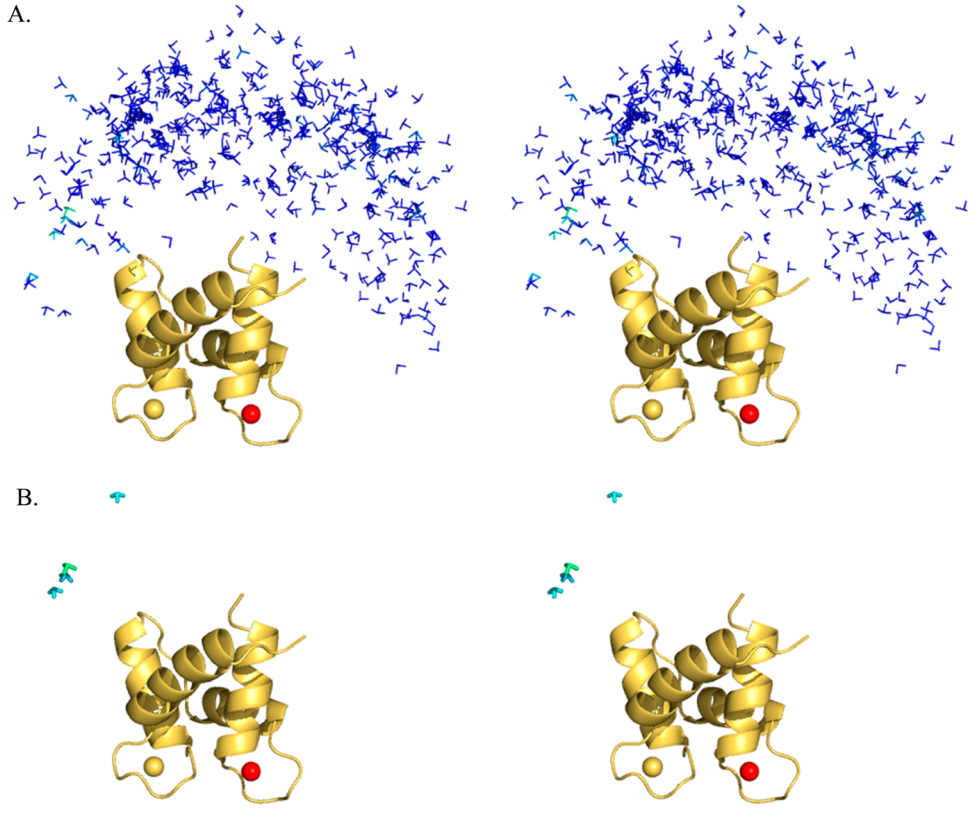


Figure 2. Stereoview (cross-eye) of (A) 400 sterically allowed conformations, color-coded according to their MaxOcc values and (B) conformations with the highest MaxOcc (ranging from 0.17 to 0.29) among these 400. Each conformation is represented as a triad of axes, centered at the center of mass of the C-terminal domain and oriented to reflect the rotation with respect to a reference structure (2K61). Color code is from 0.0 (blue) to 0.29 (green).

For the present case, only 15 RDCs (5 per metal) and around 40 PCS would actually be needed for the calculations in the absence of error. However, the use of more, seemingly redundant restraints increases the robustness of the method to the experimental uncertainty.

3.1. Simulation 1: Case of Limited Mobility. As a first example, we have evaluated how this approach performs in the presence of limited mobility. This is important to show whether any information about some residual mobility, the presence of which can be very important for protein function, can be accessed. In these cases, classical protein structure calculation methods would likely provide an average conformation of the protein, or a conformational ensemble, without being able to distinguish whether structural variability is actually needed to fulfill the experimental restraints. We have generated synthetic data starting from a compact conformation of CaM, hereon reference conformation (Figure 1), complementing it with other conformations with a different relative position of the N-terminal and C-terminal protein domains to generate an ensemble with a restricted variability in the conformational

dx.doi.org/10.1021/jp504820w | J. Phys. Chem. B 2014, 118, 10576−10587

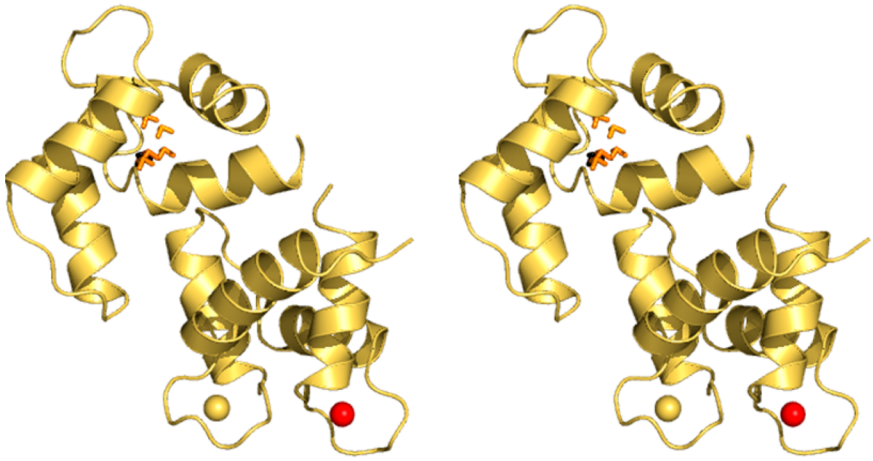


Figure 3. Stereoview (cross-eye) of conformations with the highest MaxOcc (0.66−0.71) in comparison to the reference conformation (in black).

space and averaging the PCSs and RDCs calculated for each conformation of the ensemble. The average data were perturbed with a Gaussian error (with standard deviation of 1.5 Hz for RDCs and 0.02 ppm for PCSs) and provided as "experimental" restraints. In this first example we allowed an interdomain mobility that can be described by an order parameter S_{LS} = 0.9. This means that the axial components of the anisotropy tensors derived from the RDCs of the C-terminal domain were reduced, as the result of the interdomain flexibility, by approximately 10% with respect to the axial components of the tensors determined from the N-terminal domain data, i.e., with respect to the values used in the simulation to calculate the PCSs and RDCs of each CaM conformer.

Of course, many different regions with different shapes can be defined in the conformational space which is accessible for the protein. A criterion for defining the regions mostly relevant for the calculation of the occurrence limits is to build them around the conformations with the maximum occurrence (MaxOcc). The MaxOcc is the maximum fraction of time that a single protein conformation can exist and still be compatible with the experimental observations, when taken together with any ensemble of conformations with optimized weights.[31,32,38,40−46] The analysis of the MaxOcc of single conformations has been applied to analyze the large conformational variability in CaM, free in solution or in complex[41,47] and the precollagenolytic stages of the catalytic action of matrix metalloproteinase 1.[46]

We first calculated the MaxOcc values of 400 sterically allowed conformations randomly selected over the whole conformational space of the protein.[32,38] In a previous paper, it was shown that 400 conformations are a reasonable choice when a large variability is present.[38] In the case of more limited mobility, 400 conformations are not enough as demonstrated by the evidence that only a few of those structures have MaxOcc larger than 0.2 (Figure 2). For this reason, a systematic search in the neighborhood of the conformations with highest MaxOcc was performed to select from the precalculated pool of 50 000 accessible protein conformations those with the largest MaxOcc and resulted in a cluster of structures defining a well-defined region (Figure 3 and Supporting Information), centered near the reference structure. The highest MaxOcc structures contained within this region have a MaxOcc of around 0.7. This observation raises the issue of determining the appropriate grain of the pool for the search of the high MaxOcc conformations, i.e., whether the resolution provided by a pool of 50 000 randomly generated accessible conformations is large enough to allow the reconstruction of the experimental data. To this end, Figure S2a,c shows the variation of MaxOcc values as a function of either translation of the C-terminal domain along one axis or rotation of that domain around the same axis. Each conformation in the pool was analyzed in terms of the translation of the C-terminal domain with respect to the reference structure and of the angle between the quaternions that describe the orientation of the C-terminal domain with respect to the same structure. The average distance between two nearest neighbors in the whole pool is of 3 Å and 7°: this resolution appears appropriate also in the present case of extremely limited mobility to ensure the presence of structures with the highest MaxOcc value in the pool, which can thus be recovered. In the case of no or very limited mobility, conformations in best agreement with the data can also be sought by a rigid body minimization, as previously shown by us.[41]

The above analysis indicates that MaxOcc calculations (as well as rigid body minimization in the case of very limited mobility) can accurately point out which individual conformations are in best agreement with the experimental data, although unable to fully reproduce the data. However, still no information is obtained on the size of the conformational region sampled by the protein.

To gain a deeper insight into the extent of the mobility, we calculated the maxOR of the region defined by the conformations structurally close to the highest MaxOcc conformation. The size of the smallest region in the conformational space that can reproduce the data completely (i.e., with maxOR = 1) can be regarded as a lower bound to the extent of the residual mobility.

If the five conformations with largest MaxOcc depicted in Figure 3 are collectively taken, the maxOR of such ensemble (0.84) is somewhat larger than the MaxOcc of the single conformations (0.66−0.71), but still they cannot fully represent the conformational variability of the protein. The smallest size for the region to have maxOR = 1 was chosen by gradually increasing the maximum translation and angle rotation in steps of 2 Å and 5° (see Supporting Information). This means that all conformations present in the pool and belonging to the region defined around the largest MaxOcc conformation within these

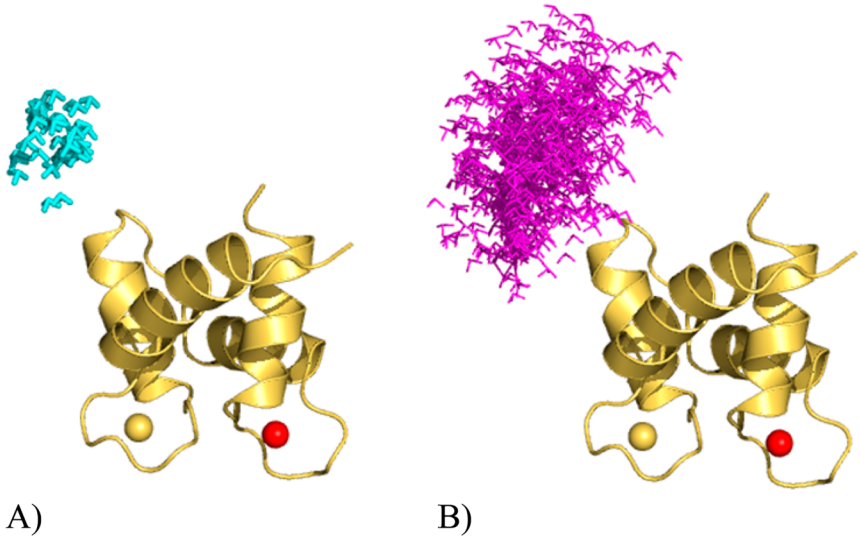


Figure 4. (A) Smallest region with maxOR = 1 and (B) smallest region with minOR = 0.54.

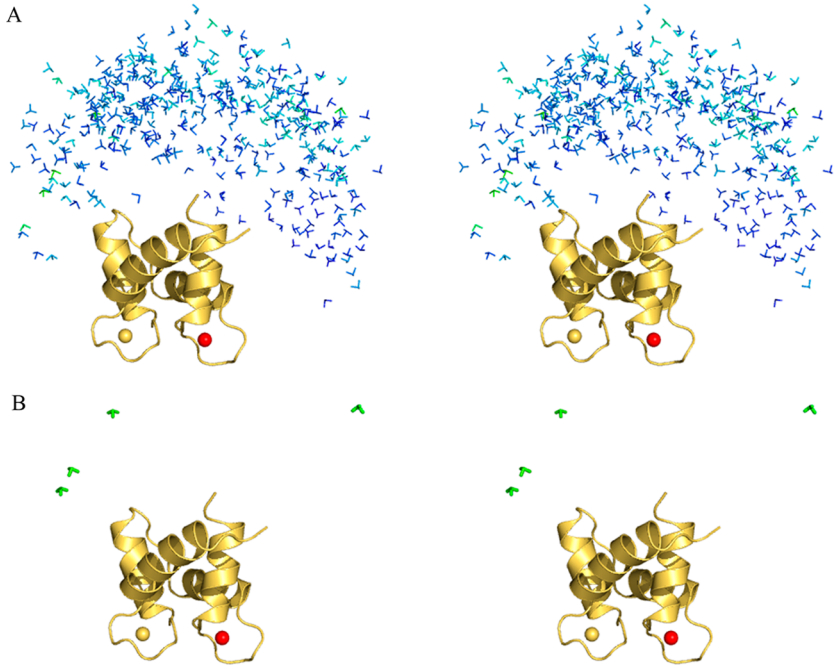


Figure 5. Stereoview (cross-eye) of (A) 400 sterically allowed conformations of CaM when bound to the DAPk1 peptide, color-coded according to their MaxOcc values, and (B) conformations with the highest MaxOcc (ranging from 0.36 to 0.43) among these 400. Color code is from <0.0 (blue) to >0.9 (red).

maximum translation and rotation limits were allowed to be freely selected and included in a conformational ensemble in full agreement with the PCS and RDC data, so as to contribute to the overall weight of the region.

The region represented in Figure 4A is able to account completely for the synthetic data; i.e., it has a maxOR of 1. It contains all the structures from the calculated pool with center of mass of the C-terminal domain translated up to 5 Å and rotated to an angle of up to 15° with respect to the structure with the highest MaxOcc (0.71). A similar result was obtained when the reference conformation used to generate the data is placed as the center of the region, the minimal size of the region that yields maxOR = 1 being defined by maximum translation and rotation of 7 Å and 10°, respectively.

By the same token, we can define the minimum occurrence of a region (minOR) as the minimum percent of time that a system must spend in a given set of conformations when included in any optimized ensemble, to allow for fitting of the experimental data. We have found that the conformations from the region spanning within a maximum translation of 13 Å and maximum rotation of 20° from the reference structure have a minOR of 0.54; i.e., the system must spend at least 54% of the time in this region to allow for fitting of the data (Figure 4B and Table S3). A large minimum occurrence for this region of the conformational space (with maxOR = 1) rules out the possibility that the corresponding structures are the accidental outcome of the motional averaging between other different conformations. The minimum occurrence of single structures, on the contrary, is always zero because any single structure can

dx.doi.org/10.1021/jp504820w | J. Phys. Chem. B 2014, 118, 10576−10587

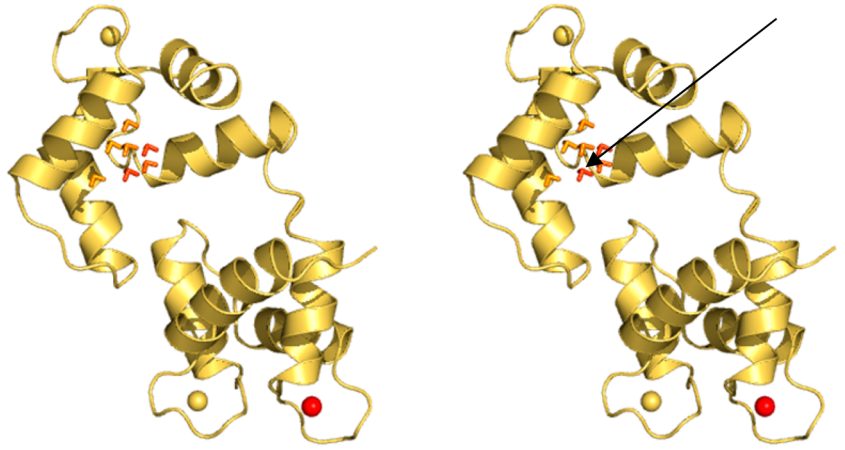


Figure 6. (A) Stereoview (cross-eye) of the conformations of CaM, when bound to the DAPk1 peptide, with the highest MaxOcc (0.65−0.81) in comparison to the experimentally determined solution structure (indicated by the arrow).

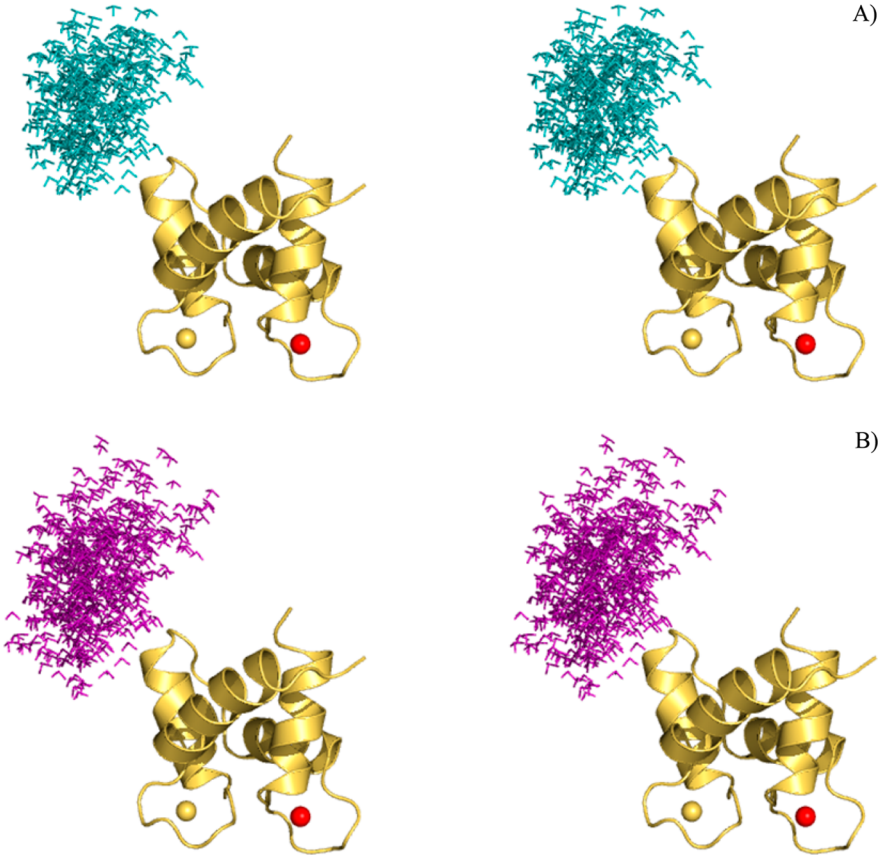


Figure 7. (A) Smallest region with maxOR = 1 and (B) smallest region with minOR = 0.10 for CaM, when bound to the DAPk1 peptide.

be excluded from the best fit ensemble and replaced by neighboring conformations without significantly affecting the quality of the fit. A minOR value of 0.54 for the identified region may seem small for a system with such a low mobility as the one under investigation, but this is intrinsic in the large degeneracy of PCS and RDC data. The minOR value of the region with maxOR = 1 shown in Figure 4A is actually equal to zero.

The conformations belonging to the maxOR = 1 region and to the minOR = 0.54 region shown in Figure 4 can be also represented in a more compact way as projections of the six-dimensional conformational space into different two-dimensional representations as described in the Supporting Information (Figures S13−S16).

3.2. Experimental Data Set: Calmodulin Bound to a Peptide Derived from DAPk1. The same approach was applied to analyze the conformational variability of CaM when bound to a peptide as the CaM-binding peptide derived from the DAPk1 protein.[34] The experimental restraints used in the calculations were the PCSs and RDCs measured by Bertini et al.[34]

The search for structures having the highest MaxOcc was first accomplished similarly to the previous case (Figure 5). A systematic search in the neighborhood of the conformations

dx.doi.org/10.1021/jp504820w | J. Phys. Chem. B 2014, 118, 10576−10587

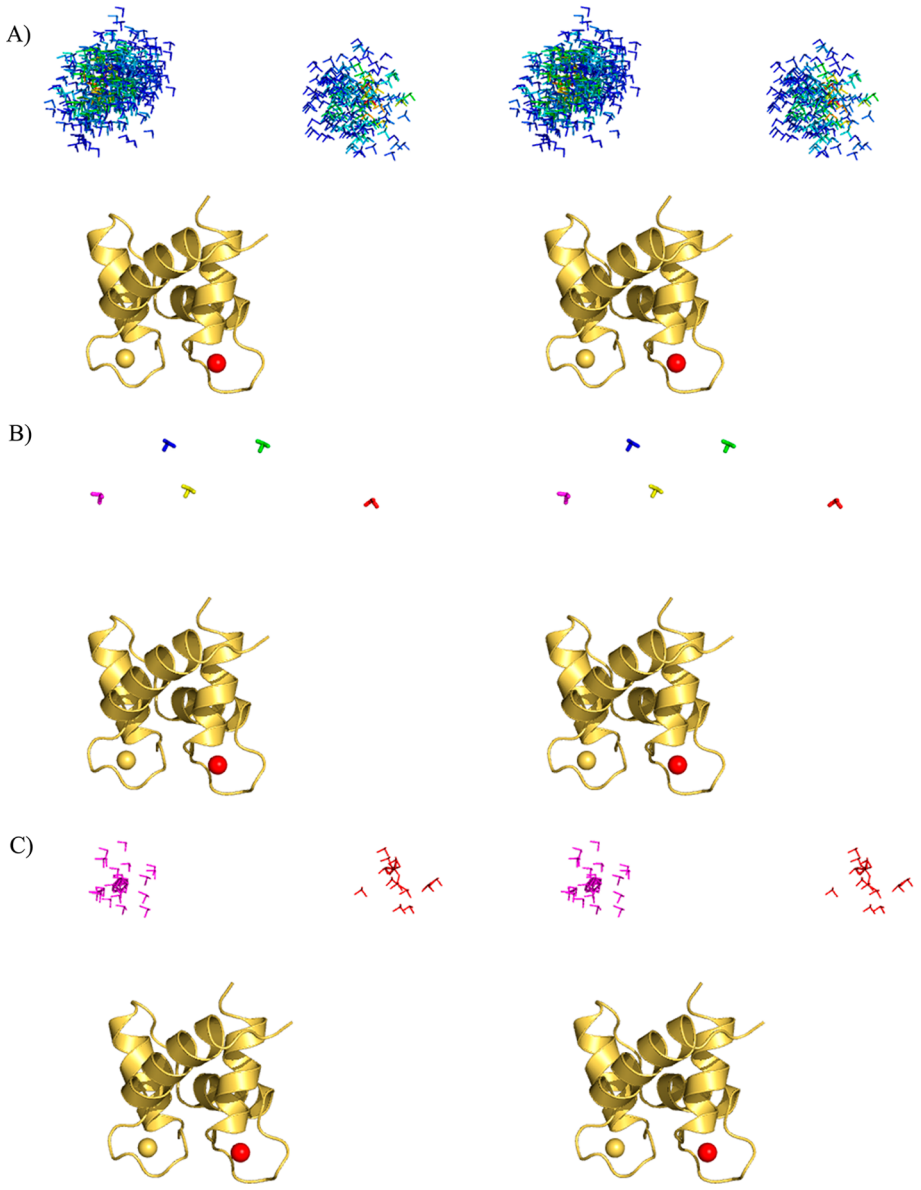


Figure 8. Stereoview (cross-eye) of (A) the regions of structures used for generating the second synthetic data set (the color indicates the relative weight, from 0.0 (blue) to >0.9 (red); the structure in the center of each region was considered with a relative weight of 1), (B) the structures having highest MaxOcc from high maximum occurrence areas, and (C) the pair of regions with the highest maxOR.

with higher MaxOcc resulted in a well-defined region (Figure 6). Interestingly, this region is centered near the solution structure.

The NMR-derived solution structure (2K61) has a MaxOcc of 0.92. This means that there might be some residual mobility so that this single structure cannot fully reproduce all experimental data. In order to understand whether the residual mobility is only due to local fluctuations of peptide bonds, we have repeated the MaxOcc calculations under the assumption that the RDCs have an order parameter S of 0.90−0.95. The MaxOcc values were found to be in general lower, rather than larger than the values calculated with $S = 1$, and correlated with those determined without considering the internal mobility. For the single case of 2K61, the inclusion of intradomain mobility with $S = 0.95$ increases MaxOcc only up to 0.94. Therefore, some (small) mobility that cannot be accounted by local fluctuations of peptide bonds can be expected.

We have then tested the X-ray structures of the protein bound to the peptide (1WRZ)[48] and to the full-length DAPk (2X0G)[49] To ensure maximal agreement of the individual domain structures to the NMR data, the NMR-refined structures of the two domains were superimposed to the structures as contained in the PDB files. Noteworthy, both structures yielded significantly lower MaxOcc (0.50 and 0.34, respectively). Although it is possible to find an ensemble consistent with the experimental data containing both such structures, there is no physical reason to preferentially include these conformations in the structural ensemble sampled by the protein in solution with respect to other conformations with larger MaxOcc.

The search for the size for the region with maxOR = 1 was chosen as described for simulation 1.

The region represented in Figure 7A is able to account completely for the experimental data. It contains all the structures from the calculated pool with translation of 11 Å and

dx.doi.org/10.1021/jp504820w | J. Phys. Chem. B 2014, 118, 10576−10587

Table 1. MaxOR (and MaxOcc in Parentheses) for the Five Regions Shown in Figure 8 and Their Pairwise Combinations[a]

	A (red)	B (green)	C (blue)	D (yellow)	E (magenta)
A	0.55 (0.54)	0.82	0.67	0.81	0.94
B		0.61 (0.57)	0.64	0.61	0.72
C			0.58 (0.55)	0.66	0.80
D				0.56 (0.55)	0.67
E					0.57 (0.54)

[a]The colors refer to the color of the conformations depicted in Figures S3g and 8.

angle of 40° with respect to the structure with highest MaxOcc. The conformations from the region spanning within a maximum translation of 15 Å and angle of 40° from the reference structure are required at least to 10% (minOR = 0.1) to allow for fitting of the data (Figure 7B).

3.3. Simulation 2: Case of Two-Site Exchange. Some proteins have been proven to exist in dynamic equilibrium between two conformationally different states, resulting from a large-scale domain rearrangement.[50,51] The presence of these two states in rapid exchange can be important to facilitate the transition to the conformation assumed in the presence of ligands. The concept of upper and lower occurrence limits, coupled together with a proper definition of the regions, allows for discovering whether a protein switches between conformations that are structurally very different, like in the case of two-site exchange. This was shown by creating a synthetic test where a two-site exchange condition is simulated. Two ensembles of conformations were generated around two randomly selected, significantly different structures (Figure 8A) and averaged PCSs and RDCs for the nuclei of the C-terminal domain were calculated (see Supporting Information). The reduction of the RDC-derived tensors for the C-terminal domain corresponded to a S_{LS} of about 0.5.

The search for structures having the highest MaxOcc was accomplished similar to the previous case. Structures belonging to the two sites, but also structures sitting in between, are found to have high MaxOcc. A simple inspection of the MaxOcc values would lead to the oversimplified conclusion that the protein samples a wide range of conformations around several centers. Five regions were then defined around the five structures with the highest MaxOcc (MaxOcc = 0.54−0.57, Table 1, diagonal values in parentheses) by selecting all conformations of the pool having a maximum of 5 Å and 10° deviation with respect to the highest MaxOcc structure in the center of the given regions. The maxOR values were then calculated for each region and are as reported in Table 1 (diagonal). Interestingly, the maxOR values for these regions were found to be not substantially different from the MaxOcc values of the single structures at the center of to the regions.

The maxOR values were then calculated over *pairs* of regions (off-diagonal elements in Table 1). While combining the central regions results in no significant increase of the maxOR, combining the two extreme regions, namely, A and E, results in a striking increase. This means that it is possible to recover a region, composed of the structures in A and E, which has by far the highest maxOR among all other regions composed of all other possible pairwise combinations (Figure 8). Comparison of Figures S3i and S3a clearly shows that by using this procedure, we succeeded in recovering the correct ensemble used to generate the data.

This demonstrates that maxOR calculations allow for recovering the conformational distribution of systems also when occurring between and around two sites, while the MaxOcc calculations for single conformations falls short. Slightly increasing the size of the A + E region (including structures with deviation up to 5 Å and 20° from the centers, i.e., up to the same maximum translation and to a maximum rotation increase of 10°) allows one to fully explain the synthetic data (i.e., the maxOR becomes equal to 1).

3.4. Simulation 3: Case of More Pronounced Mobility. We have generated a third set of synthetic data considering a region around one selected conformation, sitting exactly in the central high maximum occurrence region of the previous simulation (Figure 9A), to test whether this case can be distinguished from the previous one. The size of the selected region was chosen so as to have approximately the same reduction of the RDC-derived tensor for the C-terminal domain (i.e., $S_{LS} \approx 0.5$) as in the previous case.

The highest MaxOcc values calculated in this case are noticeably larger than in the previous simulation (up to 0.71 against 0.58), and all high MaxOcc structures span now only one relatively large but well-defined region of space. We could then select the highest MaxOcc conformation, which has a distance of 4 Å and 4° with respect to the reference conformation. Around this single conformation we have built regions of different sizes, as described for the first simulation. In this case, the smallest region in the conformational space able to reproduce completely the experimental data (having maxOR = 1) has a size defined by a translation of 7 Å and a rotation of 25° (Figure 9C). Therefore, in this case the calculations clearly pointed out that the protein spans a single compact region of the conformational space rather than two different, separated regions.

3.5. Considerations over the Minimum Occurrence. MinOR calculations were also performed for simulations 2 and 3 described above. The protocol applied in the analysis of simulation 1 provided non-negligible minimum occurrence only for regions comprising several tens of percent of the total conformational space. This implies that looking for compact regions comprising all conformations within a given distance (defined by a maximum translations and rotations from one or more centers) does not lead to any non-negligible minOR, unless the regions are so large so as to not provide meaningful information. We have thus applied a different protocol, which relies on monitoring the conformations that are used to a larger extent to fit the data. Such protocol is described in full in the Supporting Information.

This method allowed identifying for both simulations 2 and 3 regions with minOR of 0.17 and 0.21 respectively, in both cases covering less than 14% of the conformational space. These regions are, however, less straightforward to visualize in the Cartesian space (Figures S17 and S18), as the corresponding conformations were not selected on the basis of their spatial proximity. Conformations belonging to those regions are thus presented as projections of the six-dimensional conformational space into two-dimensional representations as described in the

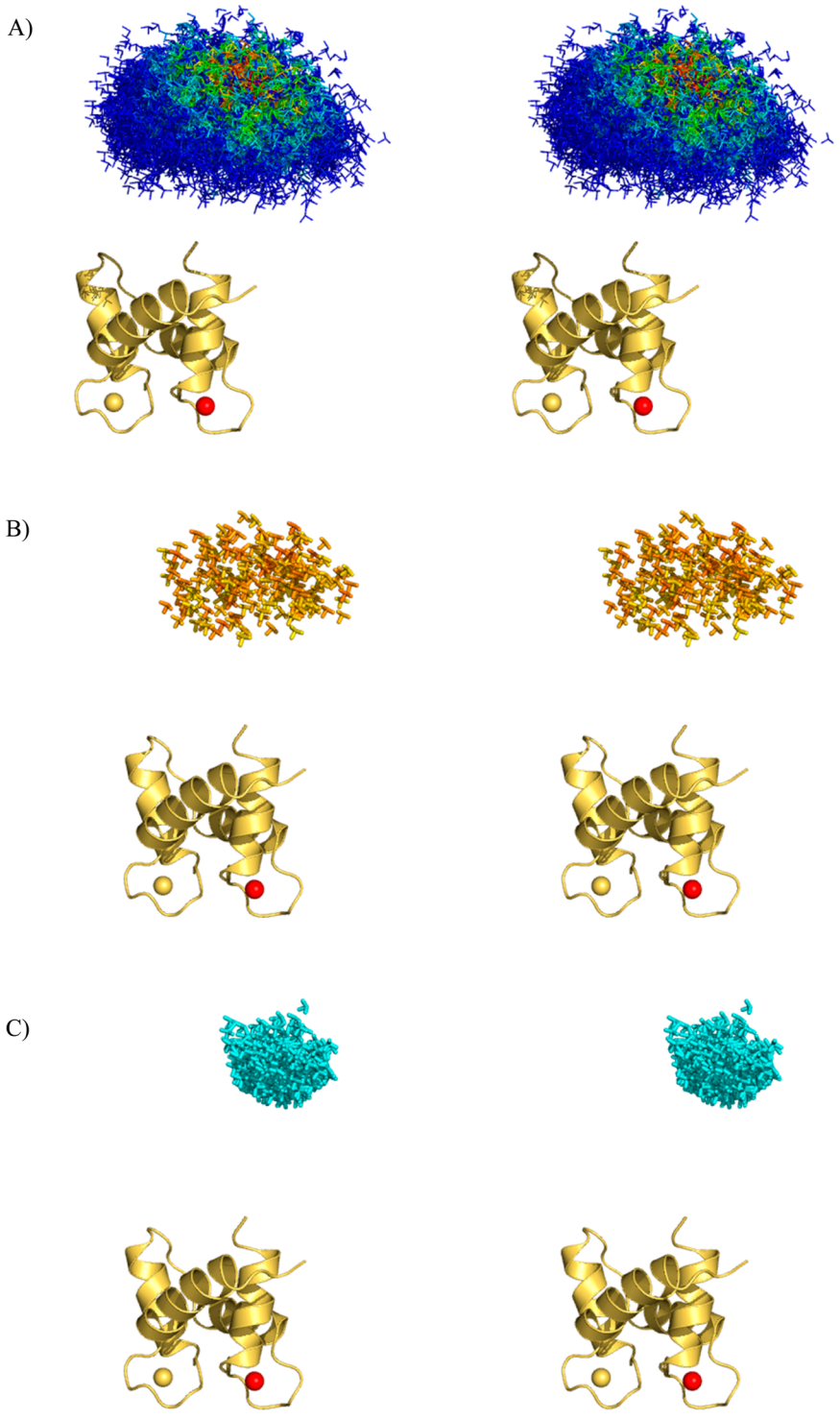


Figure 9. Stereoview (cross-eye) of (A) the region of conformations used for generating the third synthetic data set (the color indicates the relative weight, from 0.0 (blue) to >0.9 (red); the structure in the center of each region was considered with a relative weight of 1), (B) the conformations having highest MaxOcc (color code, from 0.0 (blue) to >0.9 (red)), and (C) smallest region with maxOR = 1.

Supporting Information and done also for simulation 1. The projections (Figure S5−12) show that in all cases the dominating parts of the calculated regions with maxOR = 1, as already shown, and notably also of the calculated regions with minOR equal to 0.17 or 0.21 contain the structures used in the simulations for the calculation of the synthetic data.

The low value of the minimum occurrence for the "correct" regions is to be expected in both cases, since there are several high-MaxOcc conformations in "wrong" regions of the conformational space that are "ghosts" due to the mathematical properties of the PCS and RDC functions.[41,52] However, this effect is expected to decrease with increasing number of experimental restraints: the addition of further experimental data of different nature (i.e., more independent lanthanide ions,[13,32,53] diamagnetic RDCs,[16,54] paramagnetic relaxation enhancements[44,55] but also non-NMR data such as small angle

dx.doi.org/10.1021/jp504820w | J. Phys. Chem. B 2014, 118, 10576−10587

scattering of X-rays and/or neutrons,[56,57] fluorescence resonant energy transfer,[58] or high-resolution ion mobility mass spectroscopy[59]) could improve the description of the system, decreasing the maximum allowed occurrence of the conformations and of the regions that are not actually sampled by the system and, subsequently, increasing the minimum occurrence of the region(s) that are sampled.

4. CONCLUSIONS

The presence of conformational variability in multidomain proteins and protein−protein complexes has clearly emerged in the past years to be at the basis of their function.[60,61] Depending on the latter, mobility can be restricted around a central conformation, can be so large to allow the protein to explore a large part of the conformational space, or can be restricted among few conformationally different states. Determining which of these three cases is relevant for the system under investigation is not easy. Even more difficult is determining the conformational variability sampled by the system. To address these questions, we propose the calculation of the upper and lower limits for the occurrence of regions defined in the conformational space of the protein, i.e., of ensembles of conformations.

We have tested the performance of this approach on systems with different levels of global interdomain mobility. MaxOR calculations permit in general determination of the maximum occurrence of any ensemble of conformations that the user likes to test and, as a consequence, also the minOR of all conformations excluded from such ensemble. Notably, we have also shown that suitably designed calculations of maxOR permit recovery of the conformational variability of systems switching between structures far away in the conformational space, as clearly opposed to the case of mobility around a single conformation with the same overall global order parameter. This is highly relevant in biological systems such as proteins that undergo open−closed equilibria.[50]

■ ASSOCIATED CONTENT

Supporting Information

(1) Description of the synthetic ensembles and procedure for searching the highest-scoring structures, (2) behavior of MaxOcc as a function of displacement, and (3) description of the criteria for the definition of regions. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: claudioluchinat@cerm.unifi.it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Discussions with Konstantin Berlin and David Fushman (University of Maryland), Dmitri Svergun and Maxim Petoukhov (EMBL-Hamburg), and Luca Sgheri (CNR Florence) are acknowledged. Azzurra Carlon (CERM, University of Florence, Italy) is acknowledged for the assistance in the production of the new MaxOcc software and rigid body minimization. This work was supported by Ente Cassa di Risparmio di Firenze, MIUR-FIRB Contract RBFR08WGXT, PRIN2012 SK7ASN, the EC Contracts Bio-NMR No. 261863 and BioMedBridges No. 284209, and the ESFRI Infrastructure Instruct through its core center CERM/CIRMMP. W.A. acknowledges support from the FP7-PEOPLE 2012-ITN MARIE CURIE pNMR Contract No. 317127.

■ REFERENCES

(1) Boehr, D. D.; Mcelheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313*, 1638−1642.

(2) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 954.

(3) Korzhnev, D. M.; Kay, L. E. Probing Invisible, Low-Populated States of Protein Molecules by Relaxation Dispersion NMR Spectroscopy: An Application to Protein Folding. *Acc. Chem. Res.* **2008**, *41*, 442−451.

(4) Bothe, J. R.; Nikolova, E. N.; Eichhorn, C. D.; Chugh, J.; Hansen, A. L.; Al Hashimi, H. M. Characterizing RNA Dynamics at Atomic Resolution Using Solution-State NMR Spectroscopy. *Nat. Methods* **2011**, *8*, 919−931.

(5) Dethoff, E. A.; Chugh, J.; Mustoe, A. M.; Al Hashimi, H. M. Functional Complexity and Regulation through RNA Dynamics. *Nature* **2012**, *482*, 322−330.

(6) Sicheri, F.; Kuriyan, J. Structures of Src-Family Tyrosine Kinases. *Curr. Opin. Struct. Biol.* **1997**, *7*, 777−785.

(7) Pickford, A. R.; Campbell, I. D. NMR Studies of Modular Protein Structures and Their Interactions. *Chem. Rev.* **2004**, *104*, 3557−3566.

(8) Zhang, Y.; Zuiderweg, E. R. The 70-kDa Heat Shock Protein Chaperone Nucleotide-Binding Domain in Solution Unveiled as a Molecular Machine That Can Reorient Its Functional Subdomains. *Proc. Natl. Acad. Sci. U.S.A* **2004**, *101*, 10272−10277.

(9) Tonks, N. K. Protein Tyrosine Phosphatases: From Genes, to Function, to Disease. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 833−846.

(10) Chuang, G. Y.; Mehra-Chaudhary, R.; Ngan, C. H.; Zerbe, B. S.; Kozakov, D.; Vajda, S.; Beamer, L. J. Domain Motion and Interdomain Hot Spots in a Multidomain Enzyme. *Protein Sci.* **2010**, *19*, 1662−1672.

(11) Dethoff, E. A.; Hansen, A. L.; Zhang, Q.; Al Hashimi, H. M. Variable Helix Elongation as a Tool to Modulate RNA Alignment and Motional Couplings. *J. Magn. Reson.* **2010**, *202*, 117−121.

(12) Bertini, I.; Del Bianco, C.; Gelis, I.; Katsaros, N.; Luchinat, C.; Parigi, G.; Peana, M.; Provenzani, A.; Zoroddu, M. A. Experimentally Exploring the Conformational Space Sampled by Domain Reorientation in Calmodulin. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6841−6846.

(13) Russo, L.; Maestre-Martinez, M.; Wolff, S.; Becker, S.; Griesinger, C. Interdomain Dynamics Explored by Paramagnetic NMR. *J. Am. Chem. Soc.* **2013**, *135*, 17111−17120.

(14) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. J. Determination of Domain Structure of Proteins from X-ray Solution Scattering. *Biophys. J.* **2001**, *80*, 2946−2953.

(15) Bernadò, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656−5664.

(16) Berlin, K.; Castañeda, C. A.; Schneidman-Dohovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595−16609.

(17) Huang, J.; Grzesiek, S. Ensemble Calculations of Unstructured Proteins Constrained By RDC and PRE Data: A Case Study of Urea-Denatured Ubiquitin. *J. Am. Chem. Soc.* **2010**, *132*, 694−705.

(18) Fisher, C. K.; Stultz, C. M. Constructing Ensembles for Instrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426−431.

(19) Nodet, L.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR

Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2009**, *131*, 17908−17918.
(20) Choy, W.-Y.; Forman-Kay, J. D. Calculation of Ensembles of Structures Representing the Unfolded State of an SH3 Domain. *J. Mol. Biol.* **2001**, *308*, 1011−1032.
(21) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. Constructing RNA Dynamical Ensembles by Combining MD and Motionally Decoupled NMR Rdcs: New Insights into RNA Dynamics and Adaptive Ligand Recognition. *Nucleic Acids Res.* **2009**, *37*, 3670−3679.
(22) Ryabov, Y. E.; Fushman, D. A Model of Interdomain Mobility in a Multidomain Protein. *J. Am. Chem. Soc.* **2007**, *129*, 3315−3327.
(23) Ryabov, Y. E.; Fushman, D. Analysis of Interdomain Dynamics in a Two-Domain Protein Using Residual Dipolar Couplings Together with 15N Relaxation Data. *Magn. Reson. Chem.* **2006**, *44*, S143−S151.
(24) Bashir, Q.; Volkov, A. N.; Ullmann, G. M.; Ubbink, M. Visualization of the Encounter Ensemble of the Transient Electron Transfer Complex of Cytochrome *C* and Cytochrome *C* Peroxidase. *J. Am. Chem. Soc.* **2010**, *132*, 241−247.
(25) Hulsker, R.; Baranova, M. V.; Bullerjahn, G. S.; Ubbink, M. Dynamics in the Transient Complex of Plastocyanin-Cytochrome F from *Prochlorothrix hollandica*. *J. Am. Chem. Soc.* **2008**, *130*, 1985−1991.
(26) Bonvin, A. M.; Brunger, A. T. Do NOE Distances Contain Enough Information To Assess the Relative Populations of Multi-Conformer Structures? *J. Biomol. NMR* **1996**, *7*, 72−76.
(27) Burgi, R.; Pitera, J.; Van Gunsteren, W. F. Assessing the Effect of Conformational Averaging on the Measured Values of Observables. *J. Biomol. NMR* **2001**, *19*, 305−320.
(28) Losonczi, J. A.; Andrec, M.; Fischer, M. W.; Prestegard, J. H. Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. *J. Magn. Reson.* **1999**, *138*, 334−342.
(29) Valafar, H.; Prestegard, J. H. REDCAT: A Residual Dipolar Coupling Analysis Tool. *J. Magn. Reson.* **2004**, *167*, 228−241.
(30) Shealy, P.; Simin, M.; Park, S. H.; Opella, S. J.; Valafar, H. Simultaneous Structure and Dynamics of a Membrane Protein Using REDCRAFT: Membrane-Bound Form of Pf1 Coat Protein. *J. Magn. Reson.* **2010**, *207*, 8−16.
(31) Sgheri, L. Joining RDC Data from Flexible Protein Domains. *Inverse Probl.* **2010**, *26*, 115021−115021-12.
(32) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. Conformational Space of Flexible Biological Macromolecules from Average Data. *J. Am. Chem. Soc.* **2010**, *132*, 13553−13558.
(33) Schmitz, C.; Vernon, R.; Otting, G.; Baker, D.; Huber, T. Protein Structure Determination from Pseudocontact Shifts Using ROSETTA. *J. Mol. Biol.* **2012**, *416*, 668−677.
(34) Bertini, I.; Kursula, P.; Luchinat, C.; Parigi, G.; Vahokoski, J.; Willmans, M.; Yuan, J. Accurate Solution Structures of Proteins from X-ray Data and Minimal Set of NMR Data: Calmodulin Peptide Complexes as Examples. *J. Am. Chem. Soc.* **2009**, *131*, 5134−5144.
(35) Bertini, I.; Luchinat, C.; Parigi, G. Magnetic Susceptibility in Paramagnetic NMR. *Progr. NMR Spectrosc.* **2002**, *40*, 249−273.
(36) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.
(37) Nesterov, Y. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM J. Optim.* **2012**, *22*, 341−362.
(38) Bertini, I.; Ferella, L.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Ravera, E.; Rosato, A.; Svergun, D. I. Maxocc: A Web Portal for Maximum Occurence Analysis. *J. Biomol. NMR* **2012**, *53*, 271−280.
(39) Bertini, I.; Gelis, I.; Katsaros, N.; Luchinat, C.; Provenzani, A. Tuning the Affinity for Lanthanides of Calcium Binding Proteins. *Biochemistry* **2003**, *42*, 8011−8021.
(40) Longinetti, M.; Luchinat, C.; Parigi, G.; Sgheri, L. Efficient Determination of the Most Favored Orientations of Protein Domains from Paramagnetic NMR Data. *Inverse Probl.* **2006**, *22*, 1485−1502.
(41) Bertini, I.; Gupta, Y. K.; Luchinat, C.; Parigi, G.; Peana, M.; Sgheri, L.; Yuan, J. Paramagnetism-Based NMR Restraints Provide Maximum Allowed Probabilities for the Different Conformations of Partially Independent Protein Domains. *J. Am. Chem. Soc.* **2007**, *129*, 12786−12794.
(42) Das Gupta, S.; Hu, X.; Keizers, P. H. J.; Liu, W.-M.; Luchinat, C.; Nagulapalli, M.; Overhand, M.; Parigi, G.; Sgheri, L.; Ubbink, M. Narrowing the Conformational Space Sampled by Two-Domain Proteins with Paramagnetic Probes in Both Domains. *J. Biomol. NMR* **2011**, *51*, 253−263.
(43) Luchinat, C.; Nagulapalli, M.; Parigi, G.; Sgheri, L. Maximum Occurence Analysis of Protein Conformations for Different Distributions of Paramagnetic Metal Ions within Flexible Two-Domain Proteins. *J. Magn. Reson.* **2012**, *215*, 85−93.
(44) Bertini, I.; Luchinat, C.; Nagulapalli, M.; Parigi, G.; Ravera, E. Paramagnetic Relaxation Enhancements for the Characterization of the Conformational Heterogeneity in Two-Domain Proteins. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9149−9156.
(45) Fragai, M.; Luchinat, C.; Parigi, G.; Ravera, E. Conformational Freedom of Metalloproteins Revealed by Paramagnetism-Assisted NMR. *Coord. Chem. Rev.* **2013**, *257*, 2652−2667.
(46) Cerofolini, L.; Fields, G. B.; Fragai, M.; Geraldes, C. F. G. C.; Luchinat, C.; Parigi, G.; Ravera, E.; Svergun, D. I.; Teixeira, J. M. C. Examination of Matrix Metalloproteinase-1 (MMP-1) in Solution: A Preference for the Pre-Collagenolysis State. *J. Biol. Chem.* **2013**, *288*, 30659−30671.
(47) Nagulapalli, M.; Parigi, G.; Yuan, J.; Gsponer, J.; Deraos, S.; Bamm, V. V.; Harauz, G.; Matsoukas, J.; De Planque, M.; Gerothanassis, I. P.; Babu, M. M.; Luchinat, C.; Tzakos, A. G. Recognition Pliability Is Coupled To Structural Heterogeneity: A Calmodulin-Intrinsically Disordered Binding Region Complex. *Structure* **2012**, *20*, 522−533.
(48) Kursula, P. Xdsi—A Graphical Interface for the Data Processing Program XDS. *J. Appl. Crystallogr.* **2004**, *37*, 347−348.
(49) De Diego, I.; Kuper, J.; Bakalova, N.; Kursula, P.; Wilmanns, M. Molecular Basis of the Death-Associated Protein Kinase-Calcium/Calmodulin Regulator Complex. *Sci. Signaling* **2010**, *3*, Ra6.
(50) Tang, C.; Schwieters, C. D.; Clore, G. M. Open-to-Close Transition in Apo Maltose-Binding Protein Observed by Paramagnetic NMR. *Nature* **2007**, *449*, 1078−1082.
(51) Baldwin, A. J.; Kay, L. E. NMR Spectroscopy Brings Invisible Protein States into Focus. *Nat. Chem. Biol.* **2009**, *5*, 808−814.
(52) Bertini, I.; Longinetti, M.; Luchinat, C.; Parigi, G.; Sgheri, L. Efficiency of Paramagnetism-Based Constraints To Determine the Spatial Arrangement of α-Helical Secondary Structure Elements. *J. Biomol. NMR* **2002**, *22*, 123−136.
(53) Bertini, I.; Janik, M. B. L.; Liu, G.; Luchinat, C.; Rosato, A. Solution Structure Calculations through Self-Orientation in a Magnetic Field of Cerium (III) Substituted Calcium-Binding Protein. *J. Magn. Reson.* **2001**, *148*, 23−30.
(54) Fischer, M. W.; Losonczi, J. A.; Weaver, J. L.; Prestegard, J. H. Domain Orientation and Dynamics in Multidomain Proteins from Residual Dipolar Couplings. *Biochemistry* **1999**, *38*, 9013−9022.
(55) Anthis, N. J.; Doucleff, M.; Clore, G. M. Transient, Sparsely-Populated Compact States of Apo and Calcium-Loaded Calmodulin Probed by Paramagnetic Relaxation Enhancement: Interplay of Conformational Selection and Induced Fit. *J. Am. Chem. Soc.* **2011**, *133*, 18966−18974.
(56) Petoukhov, M. V.; Svergun, D. I. Analysis of X-ray and Neutron Scattering from Biomacromolecular Solutions. *Curr. Opin. Struct. Biol.* **2007**, *17*, 562−571.
(57) Lipfert, J.; Doniach, S. Small-Angle X-ray Scattering from RNA, Proteins, and Protein Complexes. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 307−327.
(58) Ye, Y.; Blaser, G.; Horrocks, M. H.; Ruedas-Rama, M. J.; Ibrahim, S.; Zhukov, A. A.; Orte, A.; Klenerman, D.; Jackson, S. E.; Komander, D. Ubiquitin Chain Conformation Regulates Recognition and Activity of Interacting Proteins. *Nature* **2012**, *492*, 266−270.

(59) Hudgins, R. R.; Woenckhaus, J.; Jarrold, M. F. High Resolution Ion Mobility Measurements for Gas Phase Proteins: Correlation between Solution Phase and Gas Phase Conformations. *Int. J. Mass. Spec. Ion Proc.* **1997**, *165−166*, 497−507.
(60) Fragai, M.; Luchinat, C.; Parigi, G. "Four-Dimensional" Protein Structures: Examples from Metalloproteins. *Acc. Chem. Res.* **2006**, *39*, 909−917.
(61) Ravera, E.; Salmon, L.; Fragai, M.; Parigi, G.; Al-Hashimi, H.; Luchinat, C. Insights into Domain-Domain Motions in Proteins and RNA from Solution NMR. *Acc. Chem. Res.* **2014**, DOI: 10.1021/ar5002318.