J|A|C|S ARTICLES

Published on Web 09/07/2010

Conformational Space of Flexible Biological Macromolecules from Average Data

Ivano Bertini,*[,†,‡] Andrea Giachetti,[†] Claudio Luchinat,[†,‡] Giacomo Parigi,[†,‡] Maxim V. Petoukhov,[§] Roberta Pierattelli,[†,‡] Enrico Ravera,[†,‡] and Dmitri I. Svergun[§]

CERM, University of Florence, Via L. Sacconi 6, and Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy, EMBL, Hamburg Outstation, Notkestrasse 85, D-22603 Hamburg, Germany, and Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia

Received July 19, 2010; E-mail: ivanobertini@cerm.unifi.it

Abstract: The concept of *maximum occurrence* (MO), i.e., the maximum percent of time that flexible proteins can spend in any given conformation, is introduced, and a rigorous method is developed to extensively sample the conformational space and to construct MO maps from experimental data. The method is tested in a case study, the flexible two-domain protein calmodulin (CaM), using SAXS and NMR data (i.e., pseudocontact shifts and self-orientation residual dipolar couplings arising from the presence of paramagnetic lanthanide ions), revealing that the "closed" and "fully extended" conformations trapped in the crystalline forms of CaM have MOs of only 5 and 15%, respectively. Compact conformations in general have small MOs, whereas some extended conformations have MO as high as 35%, strongly suggesting these conformations to be most abundant in solution. The method is universally applicable as it requires only standard SAXS data and specific NMR data on lanthanide derivatives of the protein (using native metal sites or lanthanide tagging). The computer program is publicly available using the *grid computing* infrastructure through the authors' Web portal.

Introduction

The function of proteins is related to their structure, dynamics, and conformational flexibility, when applicable. Flexible proteins, even in the simplest generic case of two rigid domains connected by a flexible linker, may sample a wide conformational space. Their study represents a difficult task for X-ray crystallography, which may at best yield the structure of a single conformation trapped in the crystal. Solution techniques, e.g., small-angle X-ray scattering (SAXS) and nuclear magnetic resonance (NMR) spectroscopy, provide experimental observables that are averages over a manifold of conformations with different weights. The problem of recovering the protein conformational ensemble from averaged data is an *ill-defined inverse problem*[1] that admits an infinite number of solutions. The general approach adopted for globular proteins of limited mobility has been that of generating a number of conformations to better fit the experimental parameters.[2−10] The conformational freedom within protein complexes was also investigated by determining the minimum degree of mobility of the system, necessary to recover an agreement with the experimental data.[10,11] Spin labels were used to determine whether regions in the conformational space cannot be occupied by protein complexes by mapping out nuclei subject to paramagnetic relaxation enhancements, thus coming closer to the spin labels, and those showing no effects, thus farther from them.[5] Spin labels were also used to detect low-population transient conformations under equilibrium conditions using ensemble simulated annealing refinement against paramagnetic relaxation enhancement data.[6]

The concept of *maximum occurrence* (MO) of a given conformation, i.e., the maximum percent of time the system can spend in that conformation, is introduced here to help address this inverse problem from a quantitative point of view. The MO of any given conformation is calculated here as the maximum weight of this conformation at which it is still possible to reproduce the experimental data when this conformation is taken together with any number of any other conformations with variable weight. For the present study, paramagnetic NMR

† CERM, University of Florence.
‡ Department of Chemistry, University of Florence.
§ EMBL, Hamburg Outstation, and Russian Academy of Sciences.

(1) Rieping, W.; Habeck, M.; Nilges, M. *Science* **2005**, *309*, 303–306.
(2) Bertini, I.; Del Bianco, C.; Gelis, I.; Katsaros, N.; Luchinat, C.; Parigi, G.; Peana, M.; Provenzani, A.; Zoroddu, M. A. *Proc. Natl. Acad. Sci.U.S.A.* **2004**, *101*, 6841–6846.
(3) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(4) Clore, G. M.; Schwieters, C. D. *J. Mol. Biol.* **2006**, *355*, 879–886.
(5) Volkov, A. N.; Worrall, J. A. R.; Holtzmann, E.; Ubbink, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18945–18950.
(6) Iwahara, J.; Clore, G. M. *Nature* **2006**, *440*, 1227–1230.
(7) Tang, C.; Schwieters, C. D.; Clore, G. M. *Nature* **2007**, *449*, 1078–1082.
(8) Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664.
(9) Lange, O. F.; Lakomek, N.-A.; Farès, C.; Schröder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471–1475.
(10) Xu, X.; Reinle, W.; Hannemann, F.; Konarev, P. V.; Svergun, D. I.; Bernhardt, R.; Ubbink, M. *J. Am. Chem. Soc.* **2008**, *130*, 6395–6403.
(11) Xu, X.; Keizers, P. H. J.; Reinle, W.; Hannemann, F.; Bernhardt, R.; Ubbink, M. *J. Biomol. NMR* **2009**, *43*, 247–254.

spectroscopy and SAXS are used together as experimental input data for the determination of the MO of conformations corresponding to different relative positions of the two domains of calmodulin (CaM). Note, though, that the approach can be readily generalized for pairs of domains of multidomain proteins.

CaM is a two-domain protein with large conformational freedom[12,13] on which the conformations with maximum allowed probability (MAP) were obtained[14] using as experimental restraints pseudocontact shifts (pcs) and residual dipolar couplings (rdc), both generated on the C-terminal domain by different paramagnetic metal ions with magnetic anisotropy rigidly framed in the N-terminal domain. The magnetic anisotropy of each metal ion causes a specific partial alignment of the N-terminal domain, which in turn may induce a secondary alignment on the C-terminal domain depending on the conformational freedom of the latter. Likewise, pcs are observable in the C-terminal domain. Rdc and pcs measurements on CaM are available on three paramagnetic lanthanide derivatives (Tb^{3+}, Tm^{3+}, Dy^{3+}), a number that is in principle sufficient to remove ambiguities related to symmetry.[15] The protocol here developed provides the MO values throughout the conformational space rather than being limited to a search of the conformations with largest maximum probability, as performed in the previous MAP approach.

MO is not an actual probability but simply an upper limit of the fraction of time a system can spend in a given conformation. It is thus intuitive that the MO values decrease toward the actual probability if one can increase the number of experimental restraints. The small-angle scattering (SAS) techniques, both of X-rays (SAXS) and neutrons (SANS), are gaining popularity as complementary to NMR. They provide information on the shape—or average shape—of a molecule in solution.[16−18] A program, SASREF, is available for rigid-body modeling of macromolecular complexes and multidomain proteins.[19] SAXS data were previously measured for CaM[20,21] and agreed relatively well with the extended structure of CaM (PDB ID: 1UP5) observed in the solid state. More compact conformations were monitored in other SAXS studies, in agreement with molecular dynamic simulations.[22] SAS and paramagnetic NMR are used here to give reliable estimates of the MO of any conformation sampled by CaM.

Materials and Methods

Data Collection. N60D calmodulin was purchased from ProtEra srl (Florence, Italy, www.proterasrl.com). The paramagnetic NMR data (pcs and rdc for Tb^{3+}, Tm^{3+}, and Dy^{3+}) have been taken from Bertini et al.[14] In that work, CaM was dissolved in 20 mM MES and 400 mM KCl, pH 6.5.

The SAXS intensities of CaM were collected under the same buffer conditions on the X33 beamline[23] at EMBL, DESY, Hamburg. Scattering patterns were measured at several solute concentrations ranging from 1.0 to 8.0 mg/mL and processed using standard procedures.[24] The covered range of momentum transfer was $0.2 < s < 6.0\ nm^{-1}$ ($s = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and $\lambda = 0.15$ nm is the X-ray wavelength).

General Mathematical Description of the Observables. Rdc values measured on, e.g., backbone NH vectors of the protein C-terminal domain provide average information about the orientation of the C-terminal domain with respect to the magnetic susceptibility axes associated with a paramagnetic lanthanide ion bound in the N-terminal domain. The two domains are taken as rigid bodies. The measured rdc obey eq 1:[25]

$$\text{rdc (Hz)} = -\frac{1}{4\pi}\frac{B_0^{\,2}}{15kT}\frac{\gamma_I\gamma_J\hbar}{2\pi r_{IJ}^{\,3}}\Big[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\phi\Big] \quad (1)$$

where r_{IJ} is the distance between the two coupled nuclei I and J (e.g., N and ^{N}H), $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic anisotropy parameters of the magnetic susceptibility tensor of the metal, and the spherical angles θ and ϕ are those defining the orientation of the vector connecting the two coupled nuclei in the frame of the metal magnetic susceptibility tensor. One drawback of eq 1 is that multiple values of θ and ϕ give the same value of rdc, so that the rdc of the C-terminal domain are equally consistent with four symmetry-related conformations, only one being true. Collection of several sets of data using different lanthanides removes the ambiguities.[26]

The pcs provide average information on the whole conformation of the C-terminal domain. Pcs may be relatively small, because the two domains are far from one another, but they can be measured accurately. Again, pcs (eq 2) are consistent with four symmetry-related conformations,[27] and therefore, again, data on several lanthanides should be used. Pcs are described by the equation

$$\text{pcs} = \frac{1}{12\pi r_{IM}^{\,3}}\Big[\Delta\chi_{ax}(3\cos^2\vartheta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\vartheta\cos 2\varphi\Big] \quad (2)$$

where r_{IM} is the distance between the observed nucleus and the metal ion, and ϑ and φ identify the spherical coordinates of the nucleus in the frame of the metal magnetic susceptibility tensor.

SAXS provides information on the overall shape of the molecule. The experimental intensity is proportional to the scattering from a single molecule averaged over all orientations and, for any atomic structure or model, can be computed as[18]

$$I(s) = \langle|A_1(\mathbf{s}) - \rho_0 A_2(\mathbf{s}) + (\rho_b - \rho_0)A_3(\mathbf{s})|^2\rangle_\Omega \quad (3)$$

where A_i are the scattering amplitudes from the molecule in a vacuum, from the excluded volume, and from the border layer, respectively, Ω represents the solid angle in the reciprocal space,

(12) Barbato, G.; Ikura, M.; Kay, L. E.; Pastor, R. W.; Bax, A. *Biochemistry* **1992**, *31*, 5269–5278.
(13) Heidorn, D. B.; Trewhella, J. *Biochemistry* **1988**, *27*, 909–915.
(14) Bertini, I.; Gupta, Y. K.; Luchinat, C.; Parigi, G.; Peana, M.; Sgheri, L.; Yuan, J. *J. Am. Chem. Soc.* **2007**, *129*, 12786–12794.
(15) Longinetti, M.; Luchinat, C.; Parigi, G.; Sgheri, L. *Inv. Probl.* **2006**, *22*, 1485–1502.
(16) Koch, M. H.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147–227.
(17) Grishaev, A.; Wu, J.; Trewhella, J.; Bax, A. *J. Am. Chem. Soc.* **2005**, *127*, 16621–16628.
(18) Svergun, D. I.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.
(19) Petoukhov, M. V.; Svergun, D. I. *Biophys. J.* **2005**, *89*, 1237–1250.
(20) Majava, V.; Petoukhov, M. V.; Hayashi, N.; Pirilä, P.; Svergun, D. I.; Kursula, P. *BMC Struct. Biol.* **2008**, *8*, 10.
(21) Seaton, B. A.; Head, J. F.; Richardson, F. M. *Biochemistry* **1985**, *24*, 6740–6743.
(22) Likic, V. A.; Gooley, P. R.; Speed, T. P.; Strehler, E. E. *Protein Sci.* **2005**, *14*, 2955–2963.
(23) Roessle, M. W.; Klaering, R.; Ristau, U.; Robrahn, B.; Jahn, D.; Gehrmann, T.; Konarev, P. V.; Round, A.; Fiedler, S.; Hermes, S.; Svergun, D. I. *J. Appl. Crystallogr.* **2007**, *40*, s190–s194.
(24) Konarev, P. V.; Volkov, V. V.; Sokolova, A. V.; Koch, M. H. J.; Svergun, D. I. *J. Appl. Crystallogr.* **2003**, *36*, 1277–1282.
(25) Bertini, I.; Luchinat, C.; Parigi, G.; Pierattelli, R. *ChemBioChem* **2005**, *6*, 1536–1549.
(26) Bertini, I.; Kursula, P.; Luchinat, C.; Parigi, G.; Vahokoski, J.; Willmans, M.; Yuan, J. *J. Am. Chem. Soc.* **2009**, *131*, 5134–5144.
(27) Bertini, I.; Luchinat, C.; Parigi, G. *Prog. NMR Spectrosc.* **2002**, *40*, 249–273.

and **s** is the momentum transfer proportional to the scattering angle. The average scattering densities of the solvent and of the hydration shell are ρ_0 and ρ_b, respectively.

Generation of Random Structures. A pool of K random structures ($K = 56\,000$ in the present work) is generated by the program RANCH (an offspring of the ensemble generation tool from the EOM package[8]) using atomic structures of rigid domains and dummy residue (DR) representation[28] for missing linkers. The flexible linkers are generated as self-voiding random chains of DRs whereby the bond and dihedral angles are randomly taken from allowed areas of the quasi-Ramachandran plot of C^α-atoms,[29] and all the bond lengths are equal to 3.8 Å (typical distance between C^α atoms). The domain orientations are selected by random rotation around random axes taken from a quasi uniform Fibonacci grid[30] and positioned so that the N-terminal Ca connects to the last DR of the preceding linker. After generation of the pool of structures, the SAXS intensities of each of them are computed using the CRYSOL approach.[18] Experimental pcs data from the N-terminal domain are used, with FANTASIAN,[31] to obtain the magnetic anisotropy tensors of the three metals. From these, pcs and rdc are calculated for each random structure by applying eqs 1 and 2 with the program CALCALL, written for this purpose. The data sets (pcs, rdc, and SAXS) for each structure of the pool of K structures are precomputed and stored, thereby allowing a quick search through a very large (although finite) number of data sets. An example of the data sets is given in the Supporting Information, Figure S4.

Implementation of the Algorithm. To evaluate the MO of a given selected structure, an ensemble of structures extracted from the pool of K structures is exhaustively sought that is able to best reproduce the experimental data when taken together with the selected structure with a given weight. The approach consists of building one ensemble of N structures for every weight w of the considered structure, with a protocol based on simulated annealing. The target function values TF (see below) are then plotted against the weight of the considered structure, as reported in Figure 3b.

To accomplish the minimization in a fast and reliable way, the following protocol was implemented: $m = M$ families of $n < N$ structures are generated at random (all containing the selected structure at a fixed weight for which MO is being computed), and the best one in terms of TF value is selected. The "temperature" parameter T_1 used in the simulated annealing is set at the maximum value. Note that T_1 is not fixed: as the best choice for T_1 depends on the TF, it should be chosen after testing different values. For the calculations in the present work, $T_{1,\max} = 0.01$.

A random number of structures of the best family (excluding the selected structure) are exchanged with other structures from the pool: the probability of selecting one structure l depends on its weight w_l, e^{-w_l/T_2}, where T_2 is a user-defined parameter to ensure a reasonable sampling of the structures. In the present case, $T_2 = 5 \times 10^{-5}$. The TF of the present family is computed and minimized through a conjugate gradient fit of the weights of each member. The new family is accepted or rejected with a probability of $e^{-(\mathrm{TF(best)} - \mathrm{TF(new)})/T_1}$.

The exchange of structures in the best family is repeated a given number of times (50 in the present case), and then the temperature T_1 is lowered. Exchange of structures is again performed as previously done. If the minimization converges, or the lowest allowed temperature ($T_{1,\min} = 0.001$) is reached, n is increased until it reaches N, m is reduced to keep the computational time approximately constant, and the procedure starts again by setting the T_1 parameter at the maximum value and exchanging the protein structures as previously done.

For the calculations in the present work, $M = 480$ and $N = 50$. The starting value of n was set to 22. At each iteration, 4 structures were added randomly, and 11 families were removed. By numerical tests on both synthetic and real data, we found that $N = 50$ structures are optimal; by inspection of the solutions, one can easily see that only about half of the structures have non-negligible weights, thus ensuring redundancy.

The software is optimized to run on single nodes, thus ensuring perfect portability in distributed computing. Performing our calculations in the European Grid Initiative grid (http://web.eu-egi.eu), accessed through the e-NMR virtual organization (http://www.enmr.eu), we were able to compute the MO for more than 1200 structures, and of several hundred simultaneously. The flowchart of this process is summarized in the Supporting Information, Figure S3.

Definition of the Target Function TF. The target function TF that defines the discrepancy between calculated and experimental data is computed by summing up the contributions of the different restraints. For the paramagnetic NMR data ($\Delta\delta$), the overall discrepancy is computed by a quality factor q:[32]

$$q = \sum_{i=1}^{k} \frac{[\Delta\delta_i^{\mathrm{obs}} - (\sum_{j=1}^{n} w_j \Delta\delta_{ij})]^2}{\sum_{l=1}^{k} \Delta\delta_l^{\mathrm{obs}\,2}} \tag{4}$$

where k is the number of pcs/rdc and n is the number of structures in the family.

For SAXS, the discrepancy between the model and the experiment is given by a χ function:[18]

$$\chi_{\mathrm{SAXS}} = \frac{1}{k} \sum_{i=1}^{k} \frac{[I_i^{\mathrm{obs}} - cI_i^{\mathrm{calc}}]^2}{\sigma_i^{\,2}} \tag{5}$$

where I^{calc} is the average scattering intensity, calculated as

$$I^{\mathrm{calc}} = \sum_{j=1}^{n} w_j I_i^{\mathrm{calc},j} \tag{6}$$

σ_i is the standard deviation of the ith point of the scattering curve, and c is a scaling coefficient, calculated as

$$c = \left[\sum_{l=1}^{k} \frac{I_l^{\mathrm{obs}} I_l^{\mathrm{calc}}}{\sigma_l^{\,2}}\right]\left[\sum_{l=1}^{k} \frac{I_l^{\mathrm{calc}\,2}}{\sigma_l^{\,2}}\right]^{-1} \tag{7}$$

where k is the number of points in the SAXS curve. In our case, $k = 51$ (CRYSOL standard value). In order to make χ sensitive to the relative position of the two domains and not to the details of the atomic structure, SAXS data were used only up to $s = 2\ \mathrm{nm}^{-1}$, which corresponds to a spatial resolution of about 3 nm. This resolution is large enough to provide information on the relative position of the domains, as it corresponds to a sensitivity for the radius of gyration as small as 0.6 nm. This value is to be compared with the gyration radii of the single CaM domains of about 1.3 nm and with the much larger radius of gyration of the whole protein.

To allow an easier minimization, i.e., to avoid the inconveniences of the analytical form of the derivatives, the normalization of the weights of the n structures is imposed as a harmonic restraint:

(28) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. J. *Biophys. J.* **2001**, *80*, 2946–2953.
(29) Kleywegt, G. J. *J. Mol. Biol.* **1997**, *273*, 371–376.
(30) Petoukhov, M. V.; Svergun, D. I. *Biophys. J.* **2005**, *89*, 1237–1250.
(31) Banci, L.; Bertini, I.; Bren, K. L.; Cremonini, M. A.; Gray, H. B.; Luchinat, C.; Turano, P. *J. Biol. Inorg. Chem.* **1996**, *1*, 117–126.
(32) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.

$$f_{\text{weight}} = (1 - \sum_{i=1}^{n} w_i)^2 \quad (8)$$

The overall target function is given by

$$\text{TF} = a_{\text{pcs}}q_{\text{pcs}} + a_{\text{rdc}}q_{\text{rdc}} + a_{\text{SAXS}}\chi_{\text{SAXS}} + a_{\text{weight}}f_{\text{weight}} \quad (9)$$

where the weighting coefficients a_i are optimized in such a way as to have a balanced contribution by each type of restraint.

Minimization of the TF against the relative weights of the structures within each family is accomplished by analytical conjugate gradients.

Results and Discussion

Protocol for MO Calculations. To determine the MO of any given conformation, the following procedure has been developed:

(i) A representative pool of about 10^5 native-like conformations with randomized linkers is generated with an available computer program[8] (see Generation of Random Structures), and the rdc, pcs, and SAXS data corresponding to each conformation are calculated and stored (examples of data sets are reported in the Supporting Information, Figure S4). Of course, none of the data sets corresponding to a single conformation matches the experimental data if taken alone, i.e., with 100% weight.

(ii) One conformation is selected and assigned a weight lower than 100%.

(iii) A family of other conformations (typically, 50 conformations, as described in Implementation of the Algorithm) is randomly selected from the pool to complement the given conformation, and their weights are adjusted with a best-fit procedure to minimize the discrepancy with the experimental data (see Definition of the Target Function TF) when taken together with the given conformation of fixed weight.

(iv) Some members of the family are discarded and substituted with other conformations from the pool, with the selection driven by a simulated annealing protocol (see Implementation of the Algorithm), and at each step the weights of the members of the new family are again optimized. This step is repeated until convergence to a minimum value of the TF (typically 3000−3500 cycles).

(v) If the converged TF value is outside a pre-fixed tolerance, step (iv) is repeated by giving a lower weight to the selected conformation. The procedure is repeated until a weight yielding a TF value within the tolerance is obtained. At this point we have found the largest weight for that conformer that is consistent with the experimental data, i.e., its maximum occurrence, MO.

The tolerance was arbitrarily fixed to a value defined 20% higher than the lowest possible TF. This value was selected by taking into account the possible fluctuations in the results of the minimization algorithm and the error on the experimental data. The selection of a too small tolerance may thus provide incorrect results because of a non-optimal minimization. The selection of a too large tolerance, although it could somewhat increase the obtained MO values, should not change appreciably the MO ranking of the different conformations. By definition, the MO of a conformation is in fact the maximum weight that such conformation can have when considering all possible conformational ensembles and the available experimental data, so that the MO value calculated for each conformation is expected to be in any case larger than its real weight.

Such calculations can be performed simultaneously for a relatively large number (10^2−10^3) of conformations using distributed (grid) computing: actually, grid computing is ideal for this kind of approach, as the calculations are independent of one another. On the EGI grid (web.eu-egi.eu), publicly accessible through the e-NMR portal (www.enmr.eu), the calculations described above in points (iii)−(v) require an average of 6 h on a single node, and ca. 10^3 calculations can run in parallel.

MO Calculations for CaM. SAXS data on CaM solutions under the same experimental conditions as for rdc and pcs measurements[2,14] were obtained (see Supporting Information, Figure S1). The SAXS data alone were found to be in fair agreement with a single CaM structure, as obtained with the SASREF program.[19] Rdc data previously obtained on the same system, however, rule out the possibility of a single protein conformation. The rdc measured for the C-terminal domain of the protein are in fact reduced by about a factor 10 with respect to the values measured for a rigid CaM adduct and expected from the magnetic susceptibility anisotropy measured from the pcs of the N-terminal domain, where a paramagnetic metal ion is selectively substituted for the calcium(II) ion in the second binding site. Conformational averaging must thus be responsible for the observed reduction in the range of the rdc values. Therefore, rdc data show that extensive interdomain motion is present. SAXS measurements were thus used in conjunction with pcs and rdc for the calculation of MO values (see Definition of the Target Function TF).

A minimum for the TF was calculated by generating structural ensembles without any fixed conformation, equal to ∼0.22, so a tolerance of 0.266 was fixed. Most conformations provided a TF smaller than this tolerance when their weight was 0.05. Interestingly, the conformational ensembles calculated in the minimization procedures are all different, although all are equally in agreement with the experimental data. Therefore, no information can be obtained from these ensemble average calculations on the real conformational space sampled by the system, because a huge number of conformations are calculated covering an extremely large part of the conformational space, without any possibility to discriminate among them.

When the weight of the selected conformation was increased, in some cases the TF increased sizably, and in other cases it remained small. In these latter cases, many ensembles with the same low TF could be calculated, each of them containing quite different conformations with different weights. Therefore, it is immediate to define the MO of any conformation as the largest weight for which the TF is still smaller than the defined tolerance. In this way it is possible, in principle, to calculate the MO corresponding to every conformation.

The obtained results can be conveniently graphically represented using color coding, as in Figure 1, where the MO mapping is shown for two orientations (Figure 1a,b) and 400 translations exploring all the space. We can see that, with the orientation in Figure 1a, conformations on the lower right side have low MO, i.e., they can only contribute little or nothing to the observed data, while more open conformations can contribute much more. The other orientation (Figure 1b) has an overall lower MO.

One issue to address is whether a small change in the orientation with a fixed center-of-mass may provide abrupt changes in the MO. This was found not to be the case; for example, for a ±10° rotation about each of the three axes, the MO may vary up to ±2% (Figure 1c). Therefore, despite the

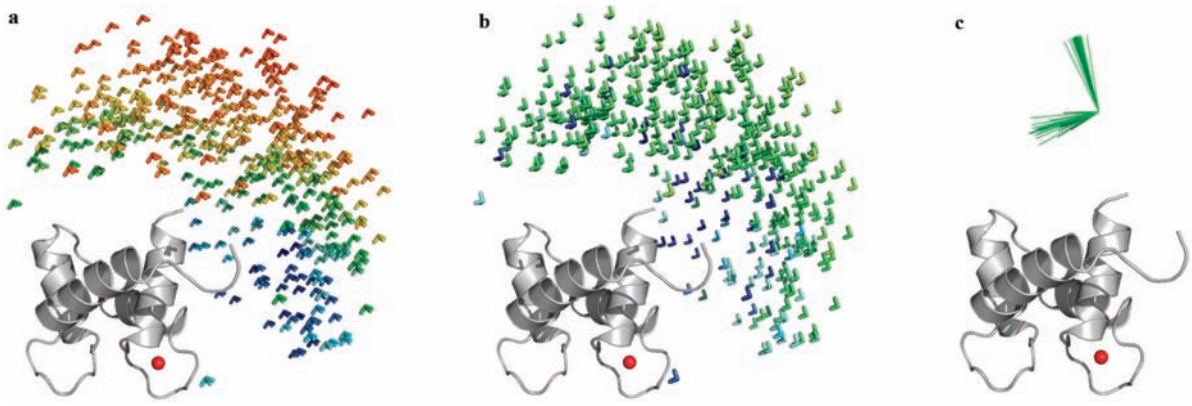

Figure 1. Orientation tensors centered in the center-of-mass of the C-terminal domain are color-coded with respect to the MO of the corresponding conformation, from blue (<5%) to red (>40%). Two different orientations (panels a and b) of the tensors are chosen to show that MO depends on both the relative domain orientation and the position. One high-MO (panel a) and one low-MO orientation (panel b) are chosen. Panel c shows the effect on MO of rotating the C-terminal domain, for a given fixed position of its center-of-mass. The C-terminal domain is rotated by ±10° about each of the three Cartesian axes.

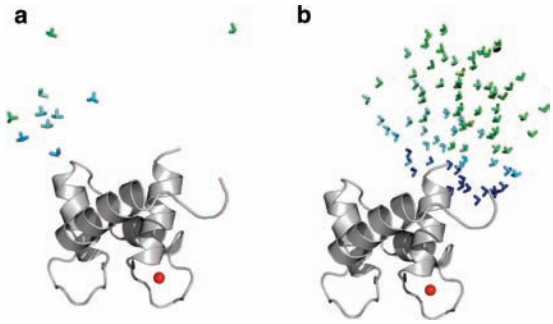

Figure 2. Orientation tensors centered in the center-of-mass of the C-terminal domain, color-coded with respect to the MO of the corresponding conformation from blue (<5%) to red (>40%), for (a) crystallographic CaM structures (PDB codes 1CLL, 1PRW, 1CDL, 1CDM, 1G4Y, 1IQ5, 1NIW, 1YR5, 2BCX, 2X0G) and (b) structures with the backbone dihedral angles of the residues in the linker region (78−81) restrained to vary within the A region of the Ramachandran plot (the starting extended 1CLL structure is indicated by a black tensor).

sensitivity of rdc to rotation, the profile of the MO map does not display sudden discontinuities.

We have also evaluated the MO for the crystal structures of free CaM[33,34] and for CaM in complexes.[26,35−40] From the calculated MO data we find that the fully elongated structure, commonly referred to as the most significant, can exist for no more than 15% of the time (Figure 2a). The closed form of free CaM can exist for at most 5% of the time, and the other closed forms observed in CaM−peptide complexes all have MO between 5 and 15% (Figure 2a).

It would be natural to ask whether some structures that are physically related to the extended one, i.e., those coming from variations of the backbone dihedral angles of the linker residues (78−81) within the same minimum in their backbone dihedral angles space (so-called A region of their Ramachandran plot[41]), can have much higher MO. Figure 2b clearly shows that this is not the case: for a number of these conformations, with the ϕ and ψ angles of the residues in the linker region (78−81) varying within the A region of the Ramachandran plot, orientation tensors centered in the center-of-mass of the C-terminal domain are shown, color-coded with respect to the MO of the corresponding conformation from blue (<5%) to red (>40%). The values of MO range from less than 5% to no more than 25%, significantly lower than the best-scoring structures (see Figure 3a). The extended 1CLL X-ray structure is indicated by a black tensor.

In general, the MO can be evaluated directly on a large number of structures, which can be obtained, for instance, with a native-like biased search using the tool RANCH of the EOM package of ATSAS.[8] Figure 3a shows the MO for 400 randomly generated structures. It confirms that the conformations having the C-terminal domain in the lower right quadrant of the frame all have low MO, while the conformations with the highest MO are clustered in the central part and in general on the outer part (more elongated) of the distribution. Figure 3b shows the TF values for all the conformations of Figure 3a as a function of their weight. It is clearly seen that there are substantial differences in the weight at which the TF value starts increasing, resulting in markedly different MO.

Simulation with Synthetic Data. In order to further demonstrate the physical meaning of the MO, a synthetic test was performed assuming that only some conformations with the ϕ and ψ values of residue 79 being restrained to vary within the A region of the Ramachandran plot and those of residue 80 being restrained to vary within the B region (black tensors in Figure 4) are possible. We then calculated the average rdc, pcs, and SAXS through eqs 1−3, introduced random errors in the synthetic data, and performed the procedure as described before. All conformations with the highest MO indeed coincide with the high-probability area occupied by the black tensors, as shown in Figure 4.

The results of this simulation show that a large MO is indeed an indication that the corresponding conformation is located in

(33) Babu, Y. S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1988**, *204*, 191–204.
(34) Fallon, J. L.; Quiocho, F. A. *Structure* **2003**, *11*, 1303–1307.
(35) Meador, W. E.; Means, A. R.; Quiocho, F. A. *Science* **1992**, *257*, 1251–1255.
(36) Meador, W. E.; Means, A. R.; Quiocho, F. A. *Science* **1993**, *262*, 1718–1721.
(37) Schumacher, M. A.; Rivard, A. F.; Bächinger, H. P.; Adelman, J. P. *Nature* **2001**, *410*, 1120–1124.
(38) Kurokawa, H.; Osawa, M.; Kurihara, H.; Katayama, N.; Tokumitsu, H.; Swindells, M. B.; Kainosho, M.; Ikura, M. *J. Mol. Biol.* **2001**, *312*, 59–68.
(39) Aoyagi, M.; Arvai, A. S.; Tainer, J. A.; Getzoff, E. D. *EMBO J.* **2003**, *22*, 766–775.
(40) Maximciuc, A. A.; Putkey, J. A.; Shamoo, Y.; MacKenzie, K. R. *Structure* **2006**, *14*, 1547–1556.
(41) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

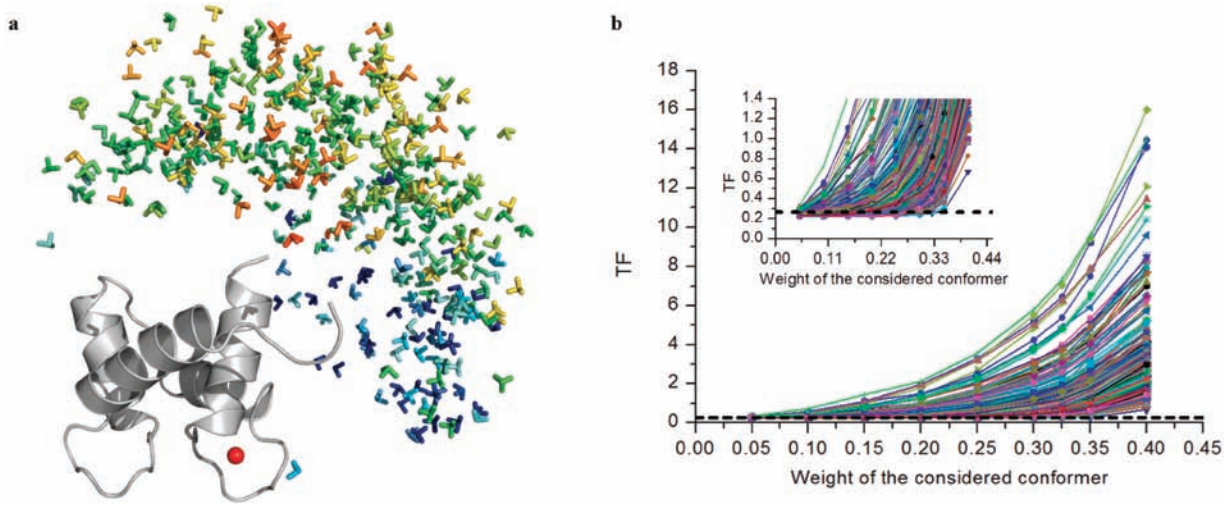

Figure 3. Orientation tensors centered in the center-of-mass of the C-terminal domain, color-coded with respect to the MO of the corresponding conformation from blue (<5%) to red (>40%) for 400 structures generated randomly with RANCH (a) and their TF vs weight curves (b). The MO is defined by the intersection between the TF curves and the predefined TF threshold (dashed line).

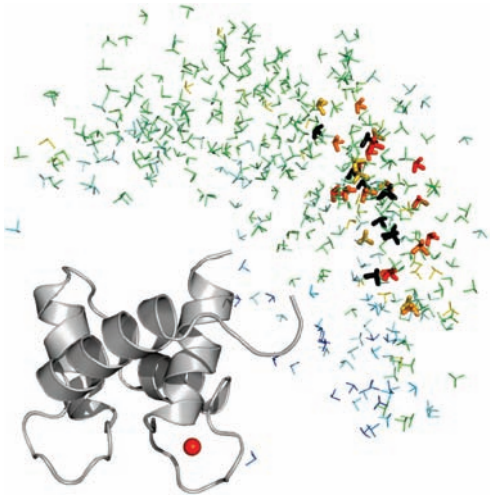

Figure 4. Comparison between the family of structures used in the simulation to generate the averaged data and the MO of the same 400 structures reported in Figure 3a. The structures are indicated through orientation tensors centered in the center-of-mass of the C-terminal domain. The tensors corresponding to the generating conformations are in black and bold, whereas those corresponding to the probe conformations are color-coded with respect to their MO from blue (<5%) to red (>40%). For clarity, those with MO >28% are bold.

a region well sampled by the system, although the MO for a single conformation does not necessarily correspond to the weight of such conformation (it actually represents its maximum allowed weight). Furthermore, almost all conformations have small but not zero MO, although they may not be sampled at all. These considerations attach further significance to the results obtained from the experimental data. Indeed, the identification of the conformations with low MO actually provides information on the regions in the protein conformational space *not* significantly sampled by the system. This is considered to be a precious piece of information,[5] especially for systems experiencing extensive motions like CaM.

Conclusions

In this paper we present a rigorous method to score any given conformation of a flexible protein according to the maximum percent of time it can exist and still be compatible with experimental data in solution. This is an evolution of a previous approach where the few conformation(s) having maximum allowed probability (MAP) were looked for.[14] Now we have the MO mapping of every structure in the conformational space. This represents a fundamental step ahead. MO mapping is computationally expensive but is possible thanks to grid-computing accessible through public portals such as www.enmr.eu.

Rdc, pcs, and SAXS data are highly complementary. Rdc provide information on the orientation. The information on the relative position of the two domains comes mainly from SAXS, while pcs contain information on both orientation and position, although their accuracy may be limited due the relatively large metal−nucleus distances. These experimental data can be obtained at any NMR infrastructures providing services, and for SAXS at synchrotrons, or for both cases even at laboratory sources.

The proposed method is general and can incorporate other data to further increase the fidelity of the MO maps. In particular, paramagnetic relaxation enhancement (PRE)[7] and any other averaged observables, which are valuable sources of structural restraints, can be readily added. Inclusion of additional types of restraints in the determination of the MO may further decrease the MO of the conformations less sampled by the system and thus help in identifying those which are actually more largely sampled. As a consequence, the MO values acquire more and more physical meaning when the number of restraints increases.

Acknowledgment. Luca Sgheri is acknowledged for discussion. This work has been supported by MIUR-FIRB contracts RBLA032ZM7, RBRN07BMCT, and RBIP06LSS2 and by the European Commission, contracts EU-NMR 026145, SPINE2-COMPLEXES 031220, and e-NMR 213010.

Supporting Information Available: Agreement between experimental and calculated data; flowchart of the code. This material is available free of charge via the Internet at http://pubs.acs.org.

JA1063923