J Biomol NMR (2012) 53:271–280
DOI 10.1007/s10858-012-9638-1

ARTICLE

MaxOcc: a web portal for maximum occurrence analysis

Ivano Bertini · Lucio Ferella · Claudio Luchinat ·
Giacomo Parigi · Maxim V. Petoukhov ·
Enrico Ravera · Antonio Rosato · Dmitri I. Svergun

Received: 2 March 2012 / Accepted: 16 May 2012 / Published online: 26 May 2012

Abstract The MaxOcc web portal is presented for the characterization of the conformational heterogeneity of two-domain proteins, through the calculation of the Maximum Occurrence that each protein conformation can have in agreement with experimental data. Whatever the real ensemble of conformations sampled by a protein, the weight of any conformation cannot exceed the calculated corresponding Maximum Occurrence value. The present portal allows users to compute these values using any combination of restraints like pseudocontact shifts, paramagnetism-based residual dipolar couplings, paramagnetic relaxation enhancements and small angle X-ray scattering profiles, given the 3D structure of the two domains as input. MaxOcc is embedded within the NMR grid services of the WeNMR project and is available via the WeNMR gateway at http://py-enmr.cerm.unifi.it/access/index/maxocc. It can be used freely upon registration to the grid with a digital certificate.

Keywords Paramagnetic NMR · Conformational heterogeneity · Ensemble averaging · Grid computing · Two-domain proteins

Introduction

Two-domain proteins may sample a wide conformational space. For these systems, solution techniques, such as nuclear magnetic resonance (NMR) and small angle scattering provide experimental observables that are weighted averages over a manifold of conformations (Bertini et al. 2004; Lindorff-Larsen et al. 2005; Iwahara and Clore 2006; Bernado et al. 2007; Xu et al. 2008; Lange et al. 2008). To recover the ensemble that originated such average observables is an ill-defined inverse problem that admits an infinite number of solutions.

We have shown how to characterize the conformational space sampled by two-domain proteins in a correct quantitative way, by the calculation of the Maximum Occurrence (MO) parameter, which is the maximum percent of time that the system can spend in a given conformation (Longinetti et al. 2006; Bertini et al. 2007; Bertini et al. 2010, 2011b). This method relies on the use of pseudocontact shifts (pcs) and self-orientation residual dipolar couplings (rdc), originating from a paramagnetic center either bound in a metal binding site or introduced by covalent tagging (Ikegami et al. 2004; Martin et al. 2007; Keizers et al. 2008; Häussinger et al. 2009; Hass et al. 2010; Su and Otting 2010; Das Gupta et al. 2011), as well as of small angle X-ray scattering (SAXS) data (Petoukhov and Svergun 2007; Bertini et al. 2009; Bernadó et al. 2010).

To sample the conformational space to an acceptable extent, it is necessary to compute MO values for hundreds of conformations. Since each conformation is analyzed

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9638-1) contains supplementary material, which is available to authorized users.

I. Bertini (✉) · L. Ferella · C. Luchinat (✉) · G. Parigi ·
E. Ravera · A. Rosato
Magnetic Resonance Center (CERM), University of Florence,
Via L. Sacconi 6, 50019 Sesto Fiorentino, FI, Italy
e-mail: ivanobertini@cerm.unifi.it

C. Luchinat
e-mail: luchinat@cerm.unifi.it

I. Bertini · C. Luchinat · G. Parigi · E. Ravera · A. Rosato
Department of Chemistry, University of Florence, Via della
Lastruccia 3, 50019 Sesto Fiorentino, FI, Italy

M. V. Petoukhov · D. I. Svergun
EMBL, Hamburg Outstation, Notkestrasse 85, 22603
Hamburg, Germany

independently of the others, the present problem can be addressed very conveniently on computational infrastructures based on distributed computing, such as a grid infrastructure. For this reason, the Maximum Occurrence approach has been originally developed using the e-NMR grid infrastructure (Bonvin et al. 2010; Loureiro-Ferreira et al. 2010), which is now managed by the WeNMR Virtual Research Community (VRC) (Wassenaar et al. 2011). However, the set up of a calculation, its execution on the grid infrastructure and the retrieval of results are all complex tasks that require significant familiarity with both the Maximum Occurrence method and the grid environment. To facilitate the use of our approach, we therefore developed a web-based portal for Maximum Occurrence calculations, called MaxOcc, which guides the user through preparing and running calculations and hides the complexity of the interaction with the grid. This model has been successfully adopted by the WeNMR VRC, which at present is the largest Life Sciences community making use of the European Grid Infrastructure. As a further support to potential MaxOcc users, we describe in this contribution various ways of using the portal and extracting information from the corresponding results.

Experimental restraints

The experimental restraints that can be presently used in MaxOcc are pcs, rdc, paramagnetic relaxation enhancements and SAXS data. Pcs data can be easily obtained as the difference in nuclear chemical shifts (in ppm) observed with the protein in the paramagnetic and diamagnetic forms. Rdc data for the H–N nuclear pairs are obtained as the difference in the doublet splitting in the indirect ^{15}N dimension in ^{1}H-^{15}N IPAP-HSQC spectra observed for the protein in the paramagnetic and diamagnetic forms. Pcs and rdc are related to the structural parameters according to the following equations (Bertini et al. 2002):

$$pcs = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\varphi\right]$$

$$rdc(Hz) = -\frac{1}{4\pi}\frac{B_0^2}{15kT}\frac{\gamma_N\gamma_H\hbar}{2\pi r_{HN}^3}\left[\Delta\chi_{ax}(3\cos^2\alpha - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\alpha\cos 2\beta\right]$$

where the symbols have the following meaning:

1. r, θ and φ are the spherical coordinates defining the position of the nuclei in the frame of the magnetic susceptibility anisotropy tensor, r_{HN} is the distance between the two coupled nuclei N and $^{\mathrm{N}}$H (set to 1.02 Å), the α and β angles define the orientation of the vector connecting the coupled N and $^{\mathrm{N}}$H nuclei in the frame of the magnetic susceptibility anisotropy tensor,
2. $\Delta\chi_{\mathrm{ax}}$ and $\Delta\chi_{\mathrm{rh}}$ are the axial and rhombic magnetic susceptibility anisotropy parameters that define the magnetic susceptibility anisotropy tensor of the paramagnetic metal ion together with the three Euler angles needed to express the protein coordinates in the frame of such tensor,
3. B_0 is the magnetic field, T the absolute temperature, k the Boltzmann constant, γ_{H} and γ_{N} the magnetogyric ratios of proton and nitrogen, respectively, and $\hbar$ the Planck constant divided by 2π.

The variables grouped in (1) can be determined from the protein structure and the orientation of the magnetic susceptibility tensor (see below). The parameters grouped in (3) are known from the experimental conditions or are known constants. Therefore, only the $\Delta\chi_{\mathrm{ax}}$ and $\Delta\chi_{\mathrm{rh}}$ values and the orientation of the magnetic susceptibility tensor (grouped in (2)) must be determined for the correct interpretation of the data. Because the structure of each protein domain is known, these parameters can be easily obtained from the best fit of the data to the structure of the domain to which the paramagnetic metal ion is bound (either directly or through a tag), according to the same equations (Bertini et al. 2002). In the presence of conformational heterogeneity, the domain to which the metal ion is not bound experiences different positions and orientations with respect to the metal-bound domain. Once the $\Delta\chi_{\mathrm{ax}}$ and $\Delta\chi_{\mathrm{rh}}$ values and the orientation of the magnetic susceptibility tensor within the metal-bound domain have been determined, the structural parameters (r, θ, φ, α and β) can be determined for any given protein conformation. This, in turn, allows the pcs and rdc values for the nuclei of the domain not bound to the metal ion to be back-calculated. The experimental data should match a weighted average of the pcs and rdc calculated for an ensemble of conformations (see Supporting Information).

^{1}H$_{\mathrm{N}}$-R_2 paramagnetic relaxation enhancements (pre's) can be determined recording HSQC (Bodenhausen and Ruben 1980) spectra with increased INEPT delay of the paramagnetic and diamagnetic samples. A two time-point measurement can be performed (Iwahara et al. 2004), in order to accurately determine pre's without any fitting procedure, according to the following equation

$$R_2^{PRE} = \frac{1}{t_b - t_a}\ln\frac{I_{dia}(t_b)I_{para}(t_a)}{I_{dia}(t_a)I_{para}(t_b)}$$

where I_{dia} and I_{para} are the peak intensities for the diamagnetic and the paramagnetic samples, respectively, at the time t_a and t_b. Pre values are proportional to the inverse of the distance to the sixth power between the paramagnetic metal ion and the observed nucleus. In the presence of mobility such distance is averaged, so that pre are provided by the following equation:

$$\begin{aligned} R_2^{PRE} &= k < r^{-6} > \\ k &= \left(\frac{\mu_0}{4\pi}\right)^2 \frac{\gamma_H^2 g_e^2 \mu_B^2 S(S+1)}{15}\left[4\tau_c + \frac{13\tau_c}{1+\omega_S^2\tau_c^2} + \frac{3\tau_c}{1+\omega_I^2\tau_c^2}\right] \\ &\quad + \frac{1}{5}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{\omega_H^2 g_e^4 \mu_B^4 S^2(S+1)^2}{(3k_B T)^2}\left[4\tau_{Curie} + \frac{3\tau_{Curie}}{1+\omega_I^2\tau_{Curie}^2}\right] \end{aligned}$$

where $\omega_\mathrm{I} = \gamma_\mathrm{I} B_0$ is the nuclear Larmor frequency, ω_S is the electron Larmor frequency ($\omega_\mathrm{S} = 658.2\ \omega_\mathrm{I}$), r is the metal-nucleus distance, g_e is the electron g factor, μ_B is the electron Bohr magneton, μ_0 is the permeability in vacuum, S is the electron spin quantum number, k_B is the Boltzmann constant, T is the temperature and τ_c is an effective correlation time which should take into account both the reorientation time of the protein and the fast motions occurring in the time scale of the overall reorientation time of the protein (Lipari and Szabo 1982; Clore et al. 1990; Brüschweiler et al. 1992; Ryabov and Fushman 2007; Iwahara and Clore 2010; Bertini et al. 2012), besides the electron relaxation time. τ_{Curie} does not depend in any case on the electron relaxation time but is an effective correlation time depending only on the global and local mobility. The value of k can be empirically determined as the time providing pre values (calculated according to the best fit ensembles) in best agreement with the calculated data (Bertini et al. 2012). In the case that the electron relaxation time is either known or negligible, because larger than the motional time constants, τ_c can be determined from amide relaxation measurements using the model free approach (Baber et al. 2001) or HYDRONMR calculations (de la Torre et al. 2000). If the electron relaxation time is larger than the motional times, the contribution from Curie relaxation (second term in the above equation) is negligible (Bertini et al. 2002).

Finally, the SAXS profile can be used as a further restraint. It provides information on the average shape of the molecule in solution (Grishaev et al. 2005; Petoukhov and Svergun 2007; Mertens and Svergun 2010). The experimental intensity related to each protein conformation is proportional to the scattering from this molecular structure averaged over all orientations, and can be computed as reported in Svergun et al. (1995). In the presence of conformational heterogeneity, the experimental profile is the weighted average of the profiles resulting from any structure in the conformational ensemble.

Maximum occurrence calculations through the MaxOcc web portal

The characterization of the conformational space of a two-domain protein using the Maximum Occurrence approach starts with the generation of a large pool (tens of thousands) of possible protein conformations, in order to allow an extensive sampling of the conformational space itself. The MO parameter is calculated for a number (a few hundred) of randomly selected conformations taken from this pool. The calculations are performed by iteratively building ensembles of conformations, in agreement with the data, each containing one of the selected conformations with a given weight. The construction of the best fitting ensembles proceeds as described in the supporting information, through the minimization of a target function quantifying the discrepancy between experimental data and data calculated from each ensemble. Ensembles are generated for each selected conformation using various fixed weights for it. The range of weights is defined by the user, e.g. based on a rough estimate of the rigidity of the system, or up to 1. The MO of each conformation is then obtained as the maximum weight for which the minimized target function does not exceed an arbitrary threshold value (Bertini et al. 2010). Figure 1 provides a pictorial illustration of the MO concept: the experimental averaged data can be reproduced using an ensemble of conformations comprising one selected conformation with a given low weight; when the weight of this selected conformation increases, a new ensemble is found, until the weight is so large that the data cannot be reproduced any more.

The MaxOcc web portal has been built using the same layout and configuration, as well as the same underlying software tools as the AMPS-NMR portal (Bertini et al. 2011a), which we previously developed also in the context of the WeNMR project to facilitate the refinement in explicit water of NMR structures using AMBER. The portal comprises two web forms, corresponding to the main steps that are needed for the calculation of the MO values (Supplementary Figure 1): (1) preparation of the input data; (2) setting of the parameters and running of the MaxOcc program. The correct format of the input files can be taken from the example files available at www.wenmr.eu/wenmr/maximum-occurrence-calculations-maxocc-web-portal. In step (1), the programs RanCh (the ensemble generation tool from the EOM package (Bernado et al. 2007)) and CALCPARA are executed. RanCh generates a very large pool of possible protein conformations by allowing the flexible residues connecting the rigid domains to change their dihedral angles in the quasi-Ramachandran space. It also computes the SAXS profile for each generated conformation. CALCPARA calculates the pcs, rdc and r^{-6} values for each conformation in the same pool. For this step the user is asked to upload/define the input for RanCh and CALCPARA (Fig. 2a). RanCh and CALCPARA calculations are run on the WeNMR grid. The status of the jobs (Scheduled, Running or Completed) can be checked in real time through a dedicated web page, via the same Jobs

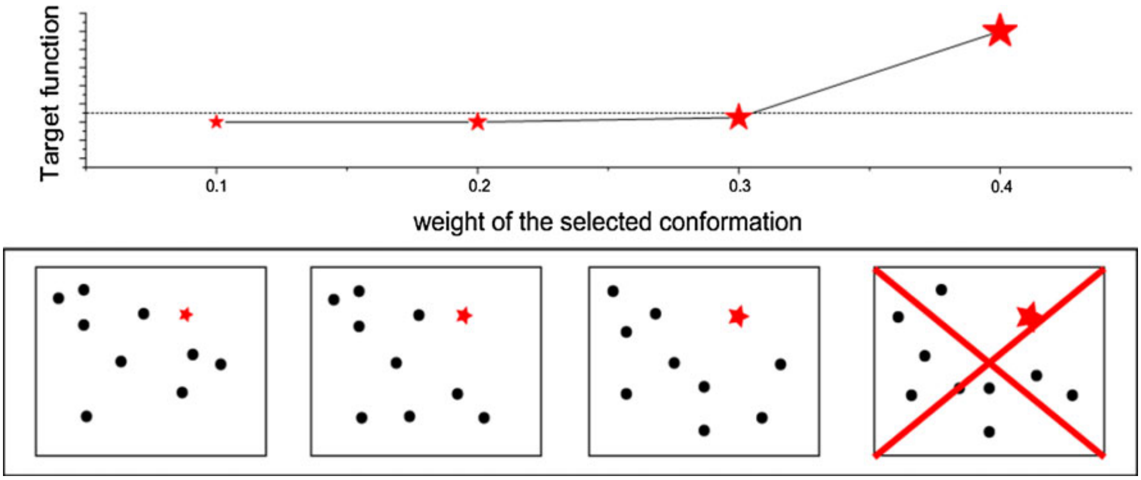


Fig. 1 Experimental data can be reproduced using different ensembles of conformations. The MO of one selected conformation (*red star*) is defined as the maximum weight that such conformation can have and still be part of an ensemble in agreement with the data (target function below the defined threshold). In the picture the *black spheres* represent the other conformations of the ensemble. When the weight of the selected conformation (proportional to the size of the *star*) is larger than its MO value, no ensemble of conformations is found to reproduce the data

management interface that we developed for the control of AMBER calculations run via the grid-enabled AMPS-NMR portal (Bertini et al. 2011a). The output files of step (1) are used to automatically build the input for the MO calculation. The web form of step (2) requires the user to specify a link to a previously performed RanCh/CALC-PARA calculation, to upload the experimental SAXS file (if available) and the constant k for the analysis of the R_2 PRE data (if available), and to define the values of the calculation parameters (Fig. 2b and Supplementary Table S1) needed to define the minimization procedure and to produce an initial analysis of the data. Standard values are prefilled in the form. MaxOcc calculations are run on the WeNMR grid, allowing a few hundred conformations to be analyzed simultaneously. In principle the optimization of the ensemble built for each of these randomly selected conformations can be run in parallel; similarly, the calculations for all the different weights analyzed for each conformation could also be run in parallel. However, this approach would require the simultaneous usage of thousands of grid nodes, which would exceed the typical availability for an individual user. Furthermore, data transfer rate can become limiting under these conditions. It is thus more practical to bundle the calculations for a few tens of different conformations/weights into a single job that will execute them consecutively. In this way, a single data transfer is needed for all of the calculations, and the number of simultaneously used cores is reduced to the order of a hundred.

After completion of all the jobs, the user can retrieve all the results of the MaxOcc analysis as a single output.

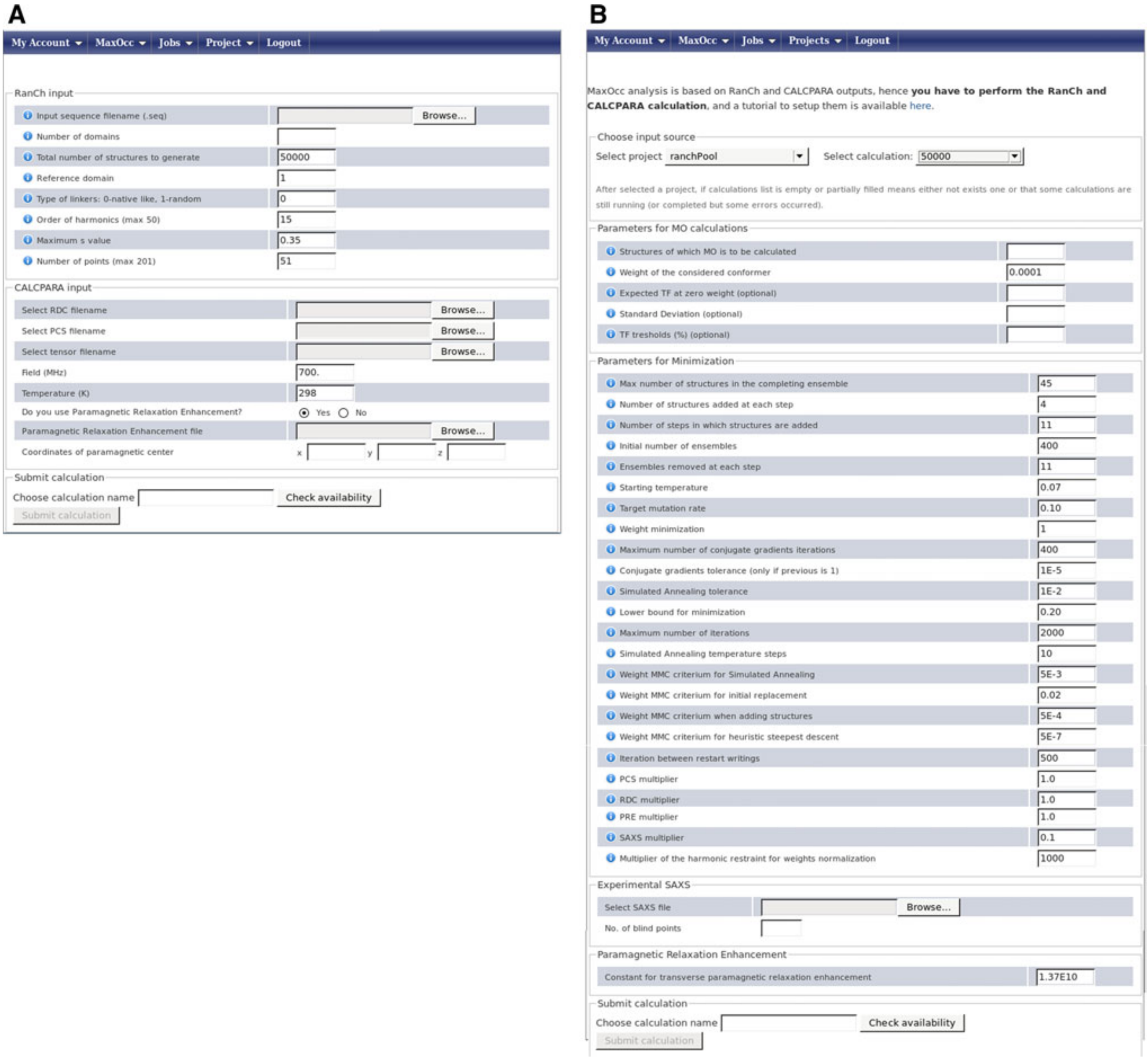
Results and discussion

In the following paragraphs, we address a selection of possible applications of the MaxOcc portal to obtain insights into the dynamic properties of calmodulin, a two-domain protein, using previously published pcs, rdc and SAXS data (Bertini et al. 2010). Pcs and rdc were measured after substitution of Tb^{3+}, Tm^{3+}, or Dy^{3+} to Ca^{2+} in the second binding site of the N-terminal domain of the protein. This selection provides both a detailed description of how to set up various types of calculations and an overview of various issues that may need to be evaluated.

Determination of the optimal number of structures in the ensemble

In order to determine the optimal number of structures that should be included in the "completing ensemble" (i.e. in the ensemble of structures that is taken together with the selected conformation) for any given dataset, multiple calculations should be performed using various numbers of structures within the ensemble. The optimal size is the smallest one for which a target function is obtained after optimization that does not decrease significantly after addition of further structures into the ensemble. A compromise between the computational time needed for the calculations and the accuracy of the conformational ensemble must be searched.

Here we exemplify how to perform this task, and at the same time demonstrate the usage of the portal, by



RanCh input

Input sequence filename (.seq)	
Number of domains	
Total number of structures to generate	50000
Reference domain	1
Type of linkers: 0-native like, 1-random	0
Order of harmonics (max 50)	15
Maximum s value	0.35
Number of points (max 201)	51

CALCPARA input

Select RDC filename	
Select PCS filename	
Select tensor filename	
Field (MHz)	700.
Temperature (K)	298
Do you use Paramagnetic Relaxation Enhancement?	● Yes ○ No
Paramagnetic Relaxation Enhancement file	
Coordinates of paramagnetic center	x y z

Submit calculation

Choose calculation name

MaxOcc analysis is based on RanCh and CALCPARA outputs, hence **you have to perform the RanCh and CALCPARA calculation**, and a tutorial to setup them is available here.

Choose input source

Select project ranchPool Select calculation: 50000

After selected a project, if calculations list is empty or partially filled means either not exists one or that some calculations are still running (or completed but some errors occurred).

Parameters for MO calculations

Structures of which MO is to be calculated	
Weight of the considered conformer	0.0001
Expected TF at zero weight (optional)	
Standard Deviation (optional)	
TF tresholds (%) (optional)	

Parameters for Minimization

Max number of structures in the completing ensemble	45
Number of structures added at each step	4
Number of steps in which structures are added	11
Initial number of ensembles	400
Ensembles removed at each step	11
Starting temperature	0.07
Target mutation rate	0.10
Weight minimization	1
Maximum number of conjugate gradients iterations	400
Conjugate gradients tolerance (only if previous is 1)	1E-5
Simulated Annealing tolerance	1E-2
Lower bound for minimization	0.20
Maximum number of iterations	2000
Simulated Annealing temperature steps	10
Weight MMC criterium for Simulated Annealing	5E-3
Weight MMC criterium for initial replacement	0.02
Weight MMC criterium when adding structures	5E-4
Weight MMC criterium for heuristic steepest descent	5E-7
Iteration between restart writings	500
PCS multiplier	1.0
RDC multiplier	1.0
PRE multiplier	1.0
SAXS multiplier	0.1
Multiplier of the harmonic restraint for weights normalization	1000

Experimental SAXS

Select SAXS file	
No. of blind points	

Paramagnetic Relaxation Enhancement

Constant for transverse paramagnetic relaxation enhancement	1.37E10

Submit calculation

Choose calculation name

Fig. 2 Screenshots of the MaxOcc web portal for step 1 (**a**) and 2 (**b**)

performing calculations with pcs, rdc and SAXS data. The dependence of the target function on the size of the completing ensemble was evaluated using a fixed weight for the selected conformation of 0.001 (Fig. 3). If the ensemble size is increased from 20 to 40 conformations, the target function decreases significantly; a slight decrease is still observed upon increasing the size up to 50–60. When the number of conformations is further raised to 70, the additional increase of the computational time is not balanced by a comparable gain in the final target function value. For the present example, an ensemble size of 50 conformations is thus preferable.

Determining the minimum number of conformations that describe the MO distribution

The MO of many different selected conformations extracted from the large pool generated by RanCh must be calculated in order to sample the conformational space available to the protein to a satisfactory extent. Increasingly flexible systems will sample a larger conformational space, thus requiring the calculation of MO values for a larger number of conformations. Consequently, the required computational effort will be higher the more dynamic a system is.

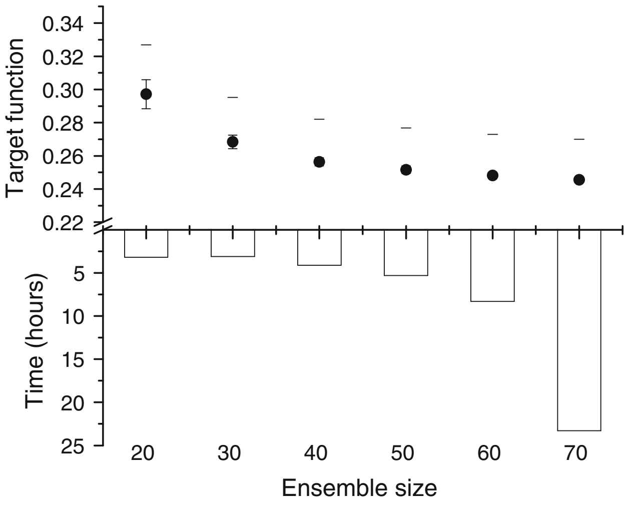


Fig. 3 The minimal target function depends on the number of conformations included in the best fit ensemble. The points indicate the minimum target function calculated by including from 20 to 70 structures in the ensemble; the *error bar* corresponds to the standard deviation of the target function value obtained from 40 calculations. Note that because of the algorithm, optimized ensembles are different even if starting from the same conformation with the same weight. The threshold used to define the MO is 10 % larger than the minimum target function value (shown by the—sign). The figure also shows the time required for completing each calculation

In order to evaluate the number of conformations that need to be characterized in the case of calmodulin, we used the portal to calculate the MO of as many as 2,000 different selected conformations. The distribution of the MO values was then compared to the corresponding distributions obtained using different subsets of this pool (Fig. 4). In all cases, a small fraction of conformations have a MO smaller than 0.1. These correspond to conformations that calmodulin cannot populate for more than 10 % of its time, otherwise it would be impossible to reproduce the experimental data. An increasingly larger number of conformations feature increasingly larger MO values, up to 0.20-0.25. The number of conformations with even larger MO values then decreases, so that a reduced fraction have a MO in the 0.25–0.30 range, and only a very small number of conformations have a MO between 0.30 and 0.40. The latter are conformations in which calmodulin can spend up to 30–40 % of its time, and thus are likely to be more important to describe calmodulin's behavior in solution. No structure has a MO larger than 0.40. Both the MO values and their variability for different groups of structures become stable at around 300 conformations, i.e. the distribution of MO values calculated from 300 selected conformations is very similar to that calculated from the entire pool. Therefore, 300 conformations can adequately represent the MO distribution within the conformational space sampled by calmodulin.

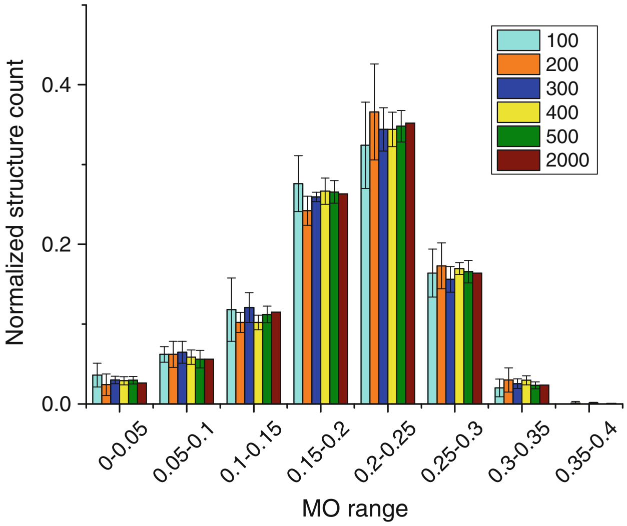


Fig. 4 Distribution of the MO values calculated for different numbers (100–2,000) of ensembles of fixed (50 conformations) size. The *error bar* is the standard deviation calculated when four different groups of structures are randomly selected

MO calculations from different datasets

The MaxOcc portal accepts as restraints paramagnetic pcs and rdc data and/or SAXS data and/or pre data. In the case of calmodulin, paramagnetic pcs and rdc restraints and SAXS restraints have been obtained under a single set of experimental conditions (Bertini et al. 2010). In order to test the relative contribution of the two sets of data (paramagnetic NMR and SAXS), MO calculations were performed by providing pcs and rdc alone, the SAXS profile alone, and all restraints together.

The MO distribution obtained from 300 randomly selected conformations in the three cases is shown in Fig. 5. The MO values of the different conformations ranged from 0 to 0.31 when all restraints were used, from 0 to 0.35 when the paramagnetic restraints only were used and from 0.23 to 0.72 when the SAXS restraints only were used. A tighter MO range with a smaller average value corresponds to an enhanced capability of the MaxOcc calculation to distinguish the conformations that can occur only for very short time lengths and the conformations that may be populated in solution for relatively long times. The MO values calculated from SAXS alone are sizably larger than the values calculated from the paramagnetic data, indicating a poorer capability to eliminate specific types of conformations within the MO approach. Nevertheless, the relative MO values correctly identify higher versus lower-probability regions (Fig. 6). When SAXS data are included together with the paramagnetic restraints, the number of conformations with the largest MO values in the distributions decreases sizably with respect to calculations performed with NMR data alone (Bertini et al. 2010).

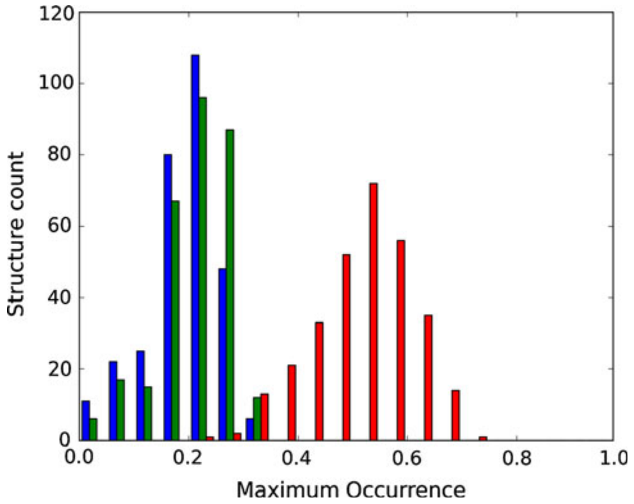


Fig. 5 Distribution of the MO values calculated for 300 conformations from SAXS restraints (*red*), pcs and rdc restraints (*green*) and both of them (*blue*)

Altogether, the present data thus indicate that SAXS by itself is less restrictive for MO calculations than NMR data by themselves, but still useful to constrain the allowed conformational space. The simultaneous use of all the data is thus advantageous.

Analysis of the best fit ensembles

A statistical analysis of the best fit ensembles is sometimes used to obtain some information on the structural variability of the investigated system (Bernado et al. 2005; Lindorff-Larsen et al. 2005; Clore and Schwieters 2006; Bernado et al. 2007; Lange et al. 2008; Gabel et al. 2008; Bertini et al. 2008). This analysis is beyond the concept of MO, for which the MaxOcc program and portal have been developed. Nevertheless, MaxOcc does calculate the conformational ensembles in best agreement with the experimental data; this prompted us to analyze whether there are common features

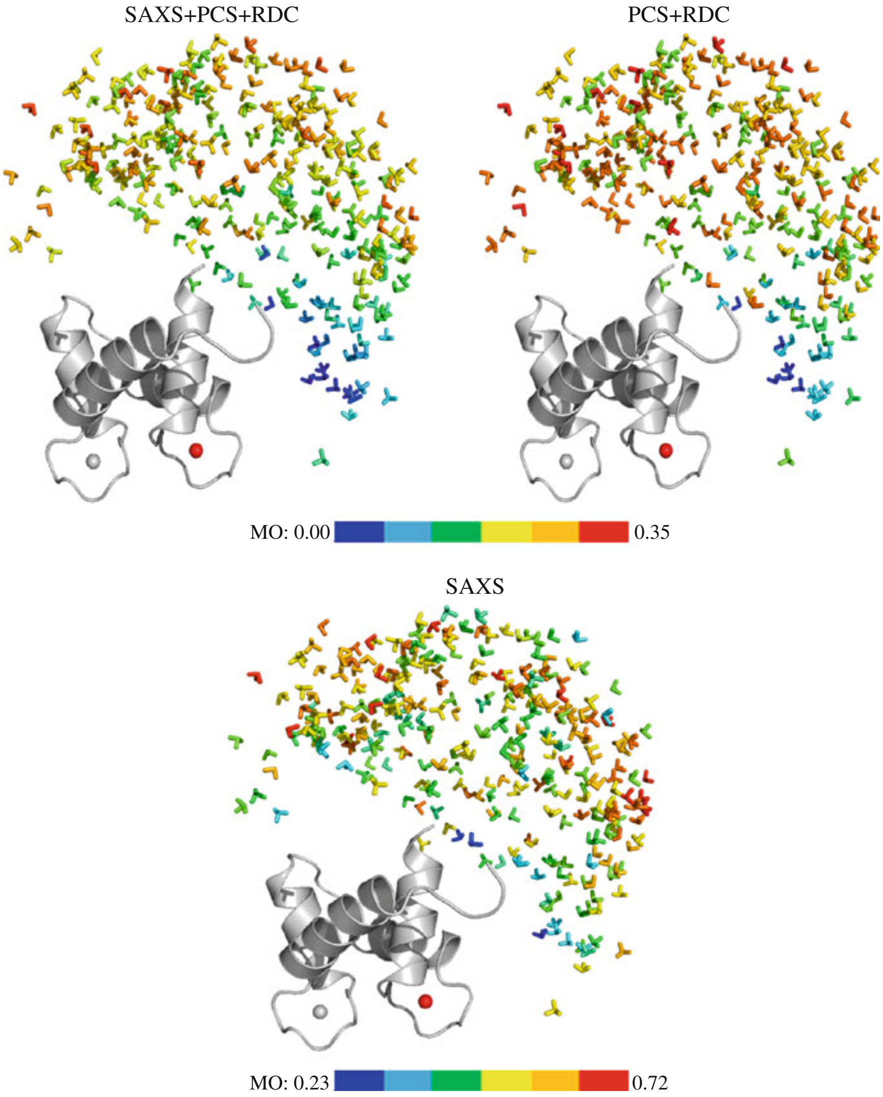


Fig. 6 Orientation tensors positioned in the centre of mass of the C-terminal domain color coded with respect to the MO of the corresponding conformation. The N-terminal domain is depicted as a cartoon, with the position of the paramagnetic Ln^{3+} ion as a *red sphere*. Each tensor represents a different position of the C-terminal domain of the protein. The tensors are placed in the center of mass of the C-terminal domain, oriented in order to reflect the orientation of the latter domain and colored from *blue* (lowest MO values) to *red* (largest MO values). The three panels depict the MO values obtained from pcs, rdc and SAXS data (range 0.00–0.35) (*left upper panel*), from pcs and rdc data (range 0.00–0.35) (*right upper panel*), and from SAXS data (range 0.23–0.72) (*bottom panel*). A script for generating the representation is available in the supporting information

in these ensembles that can be descriptive of the system. To this end, we compared the properties of the pool of 50,000 conformations generated randomly by RanCh to those observed in a set of 2,000 best fit ensembles containing 50 calmodulin conformations each (calculated for the case of a weight of the selected conformation as small as to be negligible; in this way the selected conformation practically does not contribute to the fit and thus the whole procedure becomes independent of it). The radius of gyration for the RanCh pool of conformations features a broad distribution, peaked at 20 Å and ranging from 16 to 25 Å (Fig. 7), corresponding to compact and completely extended conformations, respectively. Instead, the distribution of gyration radii of the best fit ensembles is flatter. It has comparable contributions from rather compact (18 Å) and extended (24 Å) structures, but is impoverished in structures with radius of gyration around 20 Å with respect to the RanCh-derived distribution (i.e. these structures have been rejected to an appreciable extent by the MaxOcc analysis).

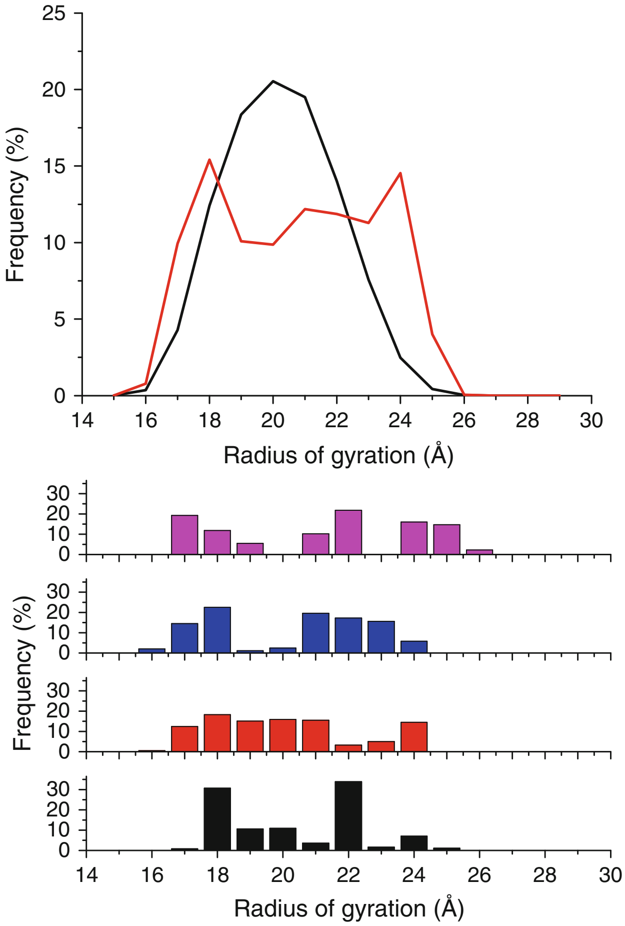


Fig. 7 Frequency of the conformations as a function of the radius of gyration in the initial pool of structures with randomized interdomain linkers (*black*) and in the selected ensembles (*red*) (*upper panel*). The latter distribution is obtained by the averaging of several calculations. The *lower panel* shows the distribution of the radius of gyration within four different ensembles, all in featuring the same level of agreement with the experimental data, in order to appreciate the distribution variability. All histograms are normalized to the integral value of unity

It should be noted that individual best-fit ensembles (any of which might actually represent the real system) differ considerably from one another (Fig. 7, lower panel). In order to devise some overall similarities between the best-fit ensembles, we clustered the conformations into regions, using the following relationship

$$\Delta t + f(1 - \cos\alpha) \leq 10$$

where Δt is the distance (in Å) between the center of the region and the center of mass of the C-terminal domain of the conformation under consideration, α is related to the orientation of the C-terminal domain with respect to the average orientation of the region, and f was set to 26. α is the angle between the quaternions corresponding to the average orientation and that under consideration. 133 regions were defined using the aforementioned procedure.

Figure 8 compares the percentage of the conformations in the 2,000 best fit ensembles belonging to the different regions with the corresponding percentage for the 50,000 conformations in the random pool. It can be immediately appreciated that MaxOcc tends to select structures in specific regions, independently of the distribution in the random pool. Anyway, even for the most populated regions there is no guarantee that they will be actually populated in solution. The lower panel in Fig. 8 shows the weight of the conformations belonging to 4 different best fit ensembles, again clustered according to the 133 regions in which we clustered the conformational space of calmodulin. It is clear that there is no region that is populated in all ensembles. These results suggest that great care should be taken when interpreting average data with ensemble approaches, as the conclusions may be significantly biased by the choice of a specific set of possible solutions.

Conclusions

We have demonstrated the use of a grid-enabled web portal to perform Maximum Occurrence (Bertini et al. 2010) calculations on a two-domain protein. The portal accepts as input data both paramagnetic NMR data (pcs, rdc and/or pre data) and SAXS profiles. These experimental data, which result from an average of the corresponding physical properties over the different conformations sampled by the protein in solution, are exploited to determine through a mathematically rigorous procedure the maximum percent of time any conformation can exist in solution.

The implementation of MaxOcc as a web portal makes its use simple to non-experts in grid computing. In addition, it also streamlines the preparation of input data and the retrieval and analysis of the calculation results, thereby

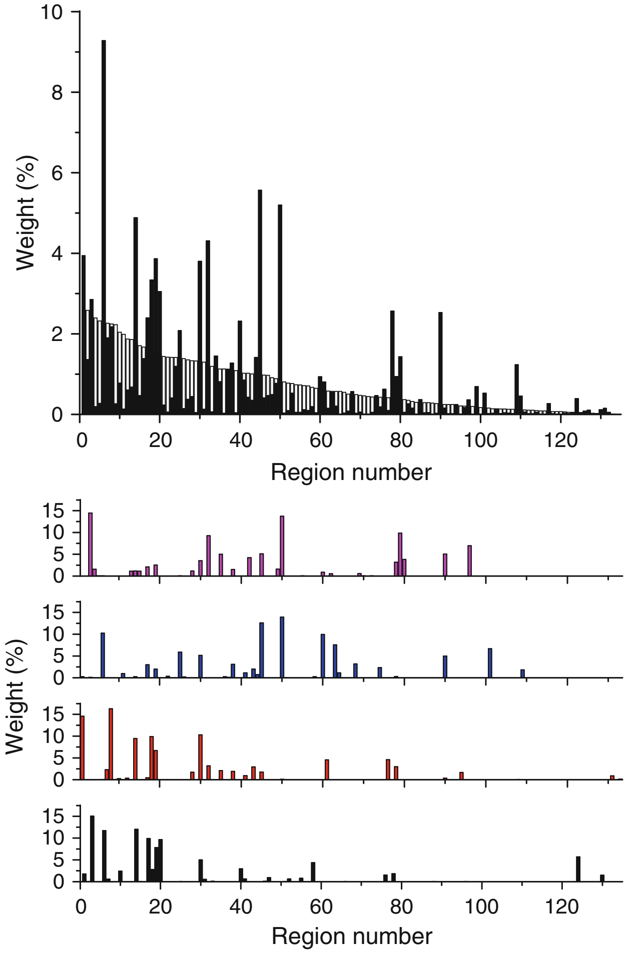


Fig. 8 Total weight of the conformations within the 2,000 best fit ensembles for the different regions of the conformational space (see text) (*black bars* in the *upper panel*). The *white bars* indicate the percentage of structures within each region of the random pool of 50,000 structures. In the *lower panel*, the weights of the conformations belonging to the different regions calculated in four best-fit ensembles are shown

significantly enhancing the user-friendliness of the approach. The results reported here warrant some caution in using conformation ensembles to infer conclusions on structural variability for two-domain proteins in solution.

Acknowledgments We acknowledge funding by MIUR-FIRB contracts RBLA032ZM7, RBRN07BMCT, and by the European Commission, contracts Bio-NMR n. 261863, East-NMR n. 228461, Biomedbridges n. 284209 and We-NMR 261572. The WeNMR project is supported by the national GRID Initiatives of Belgium, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, the United Kingdom, South Africa, Taiwan and the Latin America GRID infrastructure via the Gisela project.

References

Baber JL, Szabo A, Tjandra N (2001) Analysis of slow interdomain motion of macromolecules using NMR relaxation data. J Am Chem Soc 123:3953–3959

Bernado P, Blanchard L, Timmins P, Marion D, Ruigrok R, Blackledge M (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. Proc Natl Acad Sci USA 102:17002–17007

Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc 129:5656–5664

Bernadó P, Modig K, Grela P, Svergun DI, Tchorzewski M, Pons M, Akke M (2010) Structure and dynamics of ribosomal protein L12: An ensemble model based on SAXS and NMR relaxation. Biophys J 98:2374–2382

Bertini I, Luchinat C, Parigi G (2002) Magnetic susceptibility in paramagnetic NMR. Progr NMR Spectrosc 40:249–273

Bertini I, Del Bianco C, Gelis I, Katsaros N, Luchinat C, Parigi G, Peana M, Provenzani A, Zoroddu MA (2004) Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. Proc Natl Acad Sci USA 101:6841–6846

Bertini I, Gupta YK, Luchinat C, Parigi G, Peana M, Sgheri L, Yuan J (2007) Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. J Am Chem Soc 129:12786–12794

Bertini I, Calderone V, Fragai M, Jaiswal R, Luchinat C, Melikian M, Mylonas E, Svergun D (2008) Evidence of reciprocal reorientation of the catalytic and hemopexin-like domains of full-length MMP-12. J Am Chem Soc 130:7011–7021

Bertini I, Fragai M, Luchinat C, Melikian M, Mylonas E, Sarti N, Svergun D (2009) Interdomain flexibility in full-lenght matrix metalloproteinase-1 (MMP-1). J Biol Chem 284:12821–12828

Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E, Svergun DI (2010) Conformational space of flexible biological macromolecules from average data. J Am Chem Soc 132:13553–13558

Bertini I, Case DA, Ferella L, Giachetti A, Rosato A (2011a) A grid-enabled web portal for NMR structure refinement with AMBER. Bioinformatics 27:2384–2390

Bertini I, Luchinat C, Parigi G (2011b) Moving the frontiers in solution solid state bioNMR. A celebration of Harry Gray's 75th birthday. Coord Chem Rev 255:649–663

Bertini I, Luchinat C, Nagulapalli M, Parigi G, Ravera E (2012) Paramagnetic relaxation enhancements for the characterization of the conformational heterogeneity in two-domain proteins. Phys Chem Chem Phys (In Press). doi:10.1039/C2CP40139H

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–188

Bonvin AMJJ, Rosato A, Wassenaar T (2010) The eNMR platform for structural biology. J Struct Funct Genomics 11:1–8

Brüschweiler R, Roux B, Blackledge M, Griesinger C, Karplus M, Ernst RR (1992) Influence of rapid intramolecular motion on NMR cross-relaxation rates. A molecular dynamics study of antamanide in solution. J Am Chem Soc 114:2289–2302

Clore GM, Schwieters CD (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: a unified picture of high probability, fast motions in proteins. J Mol Biol 355:879–886

Clore GM, Szabo A, Bax A, Kay LE, Driscoll PC, Gronenborn AM (1990) Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. J Am Chem Soc 112:4989–4991

Das Gupta S, Hu X, Keizers PHJ, Liu W-M, Luchinat C, Nagulapalli M, Overhand M, Parigi G, Sgheri L, Ubbink M (2011) Narrowing the conformational space sampled by two-domain proteins with paramagnetic probes in both domains. J Biomol NMR 51:253–263

de la Torre JG, Huertas ML, Carrasco B (2000) HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. J Magn Reson 147:138–146

Gabel F, Simon B, Nilges M, Petoukhov MV, Svergun D, Sattler M (2008) A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints. J Biomol NMR 41:199–208

Grishaev A, Wu J, Trewhella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 127: 16621–16628

Hass MAS, Keizers PHJ, Blok A, Hiruma Y, Ubbink M (2010) Validation of a lanthanide tag for the analysis of protein dynamics by paramagnetic NMR spectroscopy. J Am Chem Soc 132:9952–9953

Häussinger D, Huang J, Grzesiek S (2009) DOTA-M8: an extremely rigid, high-affinity lanthanide chelating tag for PCS NMR spectroscopy. J Am Chem Soc 131:14761–14767

Ikegami T, Verdier L, Sakhaii P, Grimme S, Pescatore B, Saxena K, Fiebig KM, Griesinger C (2004) Novel techniques for weak alignment of proteins in solution using chemical tags coordinating lanthanide ions. J Biomol NMR 29:339–349

Iwahara J, Clore GM (2006) Detecting transient intermediates in macromolecular binding by paramagnetic NMR. Nature 440:1227–1230

Iwahara J, Clore GM (2010) Structure-independent analysis of the breadth of the positional distribution of disordered groups in macromolecules from order parameters for long, variable-length vectors using NMR paramagnetic relaxation enhancement. J Am Chem Soc 132:13346–13356

Iwahara J, Schwieters CD, Clore GM (2004) Ensemble approach for NMR structure refinement against H-1 paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. J Am Chem Soc 126:5879–5896

Keizers PHJ, Saragliadis A, Hiruma Y, Overhand M, Ubbink M (2008) Design, synthesis, and evaluation of a lanthanide chelating protein probe: CLaNP-5 yields predictable paramagnetic effects independent of environment. J Am Chem Soc 130:14802–14812

Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320:1471–1475

Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132

Lipari G, Szabo A (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. J Am Chem Soc 104:4546–4559

Longinetti M, Luchinat C, Parigi G, Sgheri L (2006) Efficient determination of the most favored orientations of protein domains from paramagnetic NMR data. Inv Probl 22:1485–1502

Loureiro-Ferreira N, Wassenaar TA, de Vries SJ, van Dijk M, van der Schot G, van der Zwan J, Boelens R, Giachetti A, Carotenuto D, Rosato A, Bertini I, Herrmann T, Bagaria A, Zharavin V, Jonker HR, Guentert P, Schwalbe H, Vranken WF, Dal Pra S, Mazzuccato S, Frizziero M, Traldi S, Verlato M, Bonvin AMJJ (2010) In: Proença A, Pina A, Garcia Tobio J, Ribeiro L (eds) IBERGRID 4th iberian grid infrastructure conference proceedings. Netbiblio, La Coruna, Spain, pp 360–382

Martin LJ, Hähnke MJ, Wöhnert J, Silvaggi NR, Allen KN, Schwalbe H, Imperiali B (2007) Double-lanthanide-binding tags: design, photophysical properties, and NMR applications. J Am Chem Soc 129:7106–7113

Mertens HDT, Svergun DI (2010) Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J Struct Biol 172:128–141

Petoukhov MV, Svergun DI (2007) Analysis of X-ray and neutron scattering from biomacromolecular solutions. Curr Opin Struct Biol 17:562–571

Ryabov YE, Fushman D (2007) A model of interdomain mobility in a multidomain protein. J Am Chem Soc 129:3315–3327

Su XC, Otting G (2010) Paramagnetic labelling of proteins and oligonucleotides for NMR. J Biomol NMR 46:101–112

Svergun DI, Barberato C, Koch MHJ (1995) CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773

Wassenaar T, van Dijk M, Loureiro-Ferreira N, van der Schot G, de Vries S, Schmitz C, van der Zwan J, Boelens R, Giachetti A, Ferella L, Rosato A, Bertini I, Hermann T, Jonker HR, Bagaria A, Jaravine V, Güntert P, Schwalbe H, Vranken W, Verlato M, Badoer S, Mazzuccato M, Bonvin AM, Frizziero E (2011) In: Terstyanszky G, Kiss T (eds) IWSG-life 2011: science gateway for life sciences 2011. Proceedings of the 3rd international workshop on science gateways for life sciences. London, United Kingdom

Xu X, Reinle W, Hannemann F, Konarev PV, Svergun DI, Bernhardt R, Ubbink M (2008) Dynamics in a pure encounter complex of two proteins studied by solution scattering and paramagnetic NMR spectroscopy. J Am Chem Soc 130:6395–6403