A critical assessment of methods to recover information from averaged data

Enrico Ravera,[a] Luca Sgheri[b], Giacomo Parigi,[a] Claudio Luchinat[a]*

[a]Center for Magnetic Resonance (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Via L. Sacconi 6, 50019, Sesto Fiorentino, Italy

[b]Istituto per le Applicazioni del Calcolo, Sezione di Firenze, CNR, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy

Corresponding author:

Prof. Claudio Luchinat

luchinat@cerm.unifi.it

Abstract

Conformational heterogeneity is key to the function of many biomacromolecules, but only a few groups had tried to characterize it until recently. Now, thanks to the increased throughput of experimental data and the increased computational power, the problem of the characterization of protein structural variability has become more and more popular. Several groups have devoted their efforts in trying to create quantitative, reliable and accurate protocols for extracting such information from averaged data. We analyze here the different approaches, discussing strengths and weaknesses of each. All approaches can roughly be clustered in two groups: those satisfying the maximum entropy principle and those recovering ensembles composed of a restricted number of molecular conformations. In the first case, the solution focuses on the features that are common to all the infinite solutions satisfying the experimental data; in the second case, the reconstructed ensemble shows the conformational regions where a large probability can be placed. The upper limits for the conformational probabilities (MaxOcc) can also be calculated. We also give an overview of the mainstream experimental observables, with considerations on the assumptions underlying their usage.

Introduction

Biomacromolecules are complex chemical entities. Structure is not the only key to understand their activity, as they often accomplish their complex chemical or biological tasks by more or less pronounced structural rearrangements. Thus their intrinsic mobility is a common feature that must be considered to describe biomacromolecules: it is common for a macromolecule not to be describable by a single conformation, but rather by a set of different conformations, dubbed "ensemble". This conformational variability, which is often fundamental for the function of biosystems,[1-7] is allowed for by the low energy barriers that separate the different conformations,[5] i.e.: the macromolecule can perform a sampling of the functionally relevant conformations with small differences in energy with respect to thermal energy.

In case a particular experimental observable results from a conformation-dependent feature of the biomacromolecule (e.g.: interatomic distances and/or orientations, molecular shape, etc.), one can assume that the experimental observable is given by the corresponding observable averaged over all conformations sampled by the system (see below for examples and exceptions). Based on this, possible ensembles of conformations can be determined by selecting those conformations providing averaged data in agreement with the experimental data.

However, such a reconstruction of a structural ensemble from averaged data is an ill posed inverse problem. In fact, the experimental averaged data are limited in number and do not contain enough information to determine the ensemble itself, so that the problem is severely underdetermined. An infinite number of different conformational ensembles reproducing the experimental data can actually be obtained. Therefore, no reliability can be granted to any of the reconstructed ensembles as well as to any member of the reconstructed ensembles *per se*.

Any reconstruction can be thought of as a probability distribution on the space of allowable states, be they indexed by Cartesian coordinates of atom nuclei, dihedral angles of the backbone, or Euler transformations determining the position of rigid protein domains. While there is a general agreement in what to consider an acceptable solution, i.e. a conformational ensemble in good agreement with the available experimental data, several different protocols have been developed over the years to determine possible solutions. We here recapitulate the main different protocols, discussing strengths and weaknesses of each. For examples on applications to real data, the readers can refer to available recent reviews.[8-12]

Since a unique probability distribution for the protein conformations cannot be recovered from the averaged data, the goal of any protocol is finding a solution depicting the conformations, or the conformational regions, which are certainly mostly representative to describe the conformational variability of the protein. These protocols are commonly implemented by referring to (or by different levels of combination of) two extremal paradigmatic approaches.

One extremal approach is to look for the solution which both explains the data and maintains the maximum uncertainty, and ensures that no additional assumptions on the unknown distribution is made. The solution which satisfies this criterion is the one satisfying the Maximum Entropy Principle (MEP) of Jaynes,[13] or the Kullback-Leibler divergence[14] if we assume that the solution is a perturbation on a given model maximizing the relative entropy. Even small differences in the probability distribution of the MEP solution, with respect to the unbiased distribution, are significant. This solution is often difficult to obtain because there is a large number of parameters to be fitted. This is due to the fact that loosely speaking the MEP solution is the broadest and flattest probability distribution satisfying the constraints. Solutions aiming at satisfying the MEP normally consist of a large number of individual states with small, often equal, probabilities.

The second extremal approach is to look for probability distributions where large values are concentrated in a low number of states, to point out which conformations can allow for large weights. There are different approaches with different names which can be grouped under this class: we will refer to all of them under the name of large weight (LW) solutions, i.e. solutions made by a small number of individual states with large probability.[15;16] The recovery of these states is driven by the estimate of the maximum weight, W_{max}, that can be concentrated in a given point of the space of the allowable states.[17;18] Methods looking for LW solutions tend to provide ensembles composed of states with large values of W_{max}, in order to maximize the probability values of the selected states. The solution in this case depends on the maximal number of states considered in the ensemble, and each implementation adopts a specific strategy to balance between the number of conformers and the accuracy of the solution. In many cases, the number of conformers is set to the smallest value allowing to recover the data compatibly with the experimental error: the number of structures is gradually increased until some criterion on the agreement is met. This will be discussed in more detail in the following sections.

We show here the difference between the two extremal approaches by developing the analogy used by Jaynes[13], i.e. the rolling of a die, as an example of an underdetermined problem. Suppose the only piece of information we have about a particular die is the averaged r of the points obtained by rolling the die many times. We would like to determine the probabilities p_i of obtaining the individual result i, with the obvious constrain that $\sum_{i=1}^{6} p_i = 1$. Following Jaynes, the MEP solution is given by $p_i = x^i/Z$, where x is the only real solution of the fifth-degree polynomial:

$$\sum_{i=1}^{6}(i-r)x^{i-1} = 0, \qquad (1)$$

and Z is a normalization constant determined by imposing that $\sum_{i=1}^{6} p_i = 1$. The MEP solution represents the distribution obtained with the fairest die which is compatible with the averaged *r*. Fig. 1A shows the corresponding probability distribution calculated for several values of *r*.

On the other side, using the LW approach, we look for results *i* where large probabilities are allowed by the averaged *r*. If the only available *a priori* information is the averaged value *r*, it is not possible to give any preference to any of the many possible probability distributions that can be calculated in agreement with r. However, it is possible to determine the maximum weight ($\mathrm{W}_{\mathrm{max}}$) for the result i, equal to[19;20]

$$W_{max,i} = \frac{r-j}{i-j} \qquad (2)$$

where j is the index 1 or 6 for which r is included in the interval defined by i and j. Fig. 1B shows the $\mathrm{W}_{\mathrm{max}}$ values for several values of r. The $\mathrm{W}_{\mathrm{max}}$ represents the sharp upper bound for the probability distributions satisfying the averaged data. Each single $\mathrm{W}_{\mathrm{max,i}}$ value can be obtained with a distribution satisfying the data, but in general there is no a single distribution with values $\mathrm{W}_{\mathrm{max,i}}$ for all i. Since the $\mathrm{W}_{\mathrm{max}}$ represents a bound for the probabilities, the $\mathrm{W}_{\mathrm{max,i}}$ do not sum up to 1, and a normalization would be meaningless.

Note that both the MEP solution and the $\mathrm{W}_{\mathrm{max}}$ are calculated without any assumption on both the number of states with non-zero weight out of the possible 6 states, and on the number of rolls of the die.

Suppose that the given average indicates that the solution has a strong preference towards an extremal case, i.e. r is either close to 1 or 6. Take for instance the case of a preference for the value 6, resulting in average of $r = 5.5$. In this case the MEP solution has $p_5 \cong .2238$, $p_6 \cong .6637$ and smaller p_i for i<5, and the $\mathrm{W}_{\mathrm{max}}$ estimate provides $W_{max,5} = .5$, $W_{max,6} = .9$, and smaller $\mathrm{W}_{\mathrm{max}}$ values for the results 1 to 4. Both the MEP solution and the $\mathrm{W}_{\mathrm{max}}$ estimates thus point out the preferred state.

In the LW approach, we can reconstruct the averaged $r = 5.5$ by averaging the possible values of the die. We can implement a selection procedure where the number of states (the different values of the die) with non-zero probability is constrained to a fixed value. Of course, we have no solution by assuming a single state with non-zero probability. Raising the number of states with non-zero probabilities to two is sufficient to reconstruct the average. In this case, we have the unique solution

$p_5 = p_6 = 0.5$ which recovers perfectly the average (Fig. 2). Adding further states with non-zero weight does not improve the agreement (it is indeed impossible to recover the data with three states, a quirk due to the sparsity of the allowable observables). Using four rolls of the die (i.e. composing the probability distribution by summing up four $p = 0.25$ terms), besides the solution $p_5 = p_6 = 0.5$, there is also the solution $p_6 = 0.75$ and $p_4 = 0.25$. If the number of rolls is raised further, also the other states may have a probability larger than 0, but anyway $p_6 \geq \Sigma_{i=1}^{5} p_i$. If the number of rolls is 10 and $p_1 = 0.1$ (the die provided 1 only once), p_6 must be 0.9 in order to obtain $r = 5.5$. This value of p_6 corresponds to its $\mathrm{W}_{\mathrm{max}}$ value.

Suppose on the other hand that the given average coincides with the fair die case $r = 3.5$. The MEP solution is then the uniform probability distribution $p_i = 1/6$, while the $\mathrm{W}_{\mathrm{max}}$ is maximal in $W_{max,3} = W_{max,4} = 5/6$ and becomes minimal in $W_{max,1} = W_{max,6} = 1/2$ (see Fig. 1). In this case the differences between the two approaches are more evident: the MEP approach indeed recovers the solution expected from a fair die, whereas the $\mathrm{W}_{\mathrm{max}}$ values indicate a preference towards the values 3 and 4. The MEP solution warns us that there is no reason to prefer one result or the other, since the uniform distribution explains the data. On the other hand if one assumes - or has external evidence - that there is anyway an asymmetry in the distribution (loaded die), the LW solutions point out where states with largest probabilities can be placed. The selection procedure introduced above can recover exact solutions with two states: these solutions are not unique, since, for instance, with $p_j = p_{7-j} = 0.5$, for j equal to 1, 2 or 3, the averaged $r = 3.5$ is anyway obtained. Even in these simple examples the selection procedure is confronted with the under-determined nature of the problem, being able to provide different solutions with the same agreement with the data.

Several ensemble averaging protocols have been developed employing different criteria to select among solutions with a similar agreement with the data. The chosen criterion clearly influences the recovered solution. Most protocols are implemented with a compromise between the LW and the MEP approaches: we will refer to LW or MEP methods according to the privileged focus.

Practically, LW solutions are often recovered by searching for ensembles composed of a restricted number of molecular conformations providing averaged data in agreement with the experimental data. The solution recovered with the minimum number of states represents the “simplest” solution allowed for by the averaged data, and shows the conformational regions where a large probability can be placed. It is not meant to be the real solution, but the one better representing the information content of the data, because describing few key conformations needed to reconstruct the averaged data.

On the other hand, MEP methods focus on determining the features common to all the infinite solutions satisfying the experimental data. In order to represent all the solutions, a large number of states should be allowed in the ensemble. The asymmetry of the MEP solution reveals the preference of the system for specific conformations. Every solution must thus show this preference, even if not to the extent of the MEP solution. It is not meant to be the real solution, but the one better representing the minimal set of features common to all solutions.

Experimental observables

The first question we should ask ourselves is whether the analysis of the experimental data can provide evidence of the occurrence of conformational heterogeneity. To answer this question with sufficient detail, we must consider the physical picture of the way conformational changes induce the averaging in the different experimental methodologies. The mostly used experimental data are those providing long-range structural information or those depending on the overall shape of the molecule, as can be obtained from NMR,[8-10;21] both paramagnetic[11;16;22] (pseudocontact shifts –PCSs–, self-orientation residual dipolar couplings –pRDCs–, paramagnetic relaxation enhancements

–PREs–) and diamagnetic (residual dipolar couplings –RDCs–, relaxation measurements),[23-26] optical measurements (fluorescence resonant energy transfer –FRET–),[27;28] EPR (double electron-electron resonance –DEER–),[29-31] small angle scattering (SAXS and SANS).[32;33]

For some of these methodologies, the physical picture for the averaging over multiple conformations is rather simple, as the corresponding observables depend only on features that a particular atom or moiety of the biomolecule have at a given time. These observables are, for instance, SAXS, SANS and DEER, where the detection is instantaneous with respect to the lifetime of the conformation, and PCSs and pRDCs, when the rate of interconversion between conformations is larger than the differences in nuclear chemical shifts among the different conformations. In these cases, the experimental data is the average of the observables calculated for all conformations sampled by the system.

Also diamagnetic RDCs are commonly analyzed as average of the data calculated from individual molecular conformations, under the assumption that they are rigid during the time course of the interaction of the molecule with the alignment medium. This assumption might actually fail in representing the real physical picture, as these interactions may perturb the molecular conformation, thus questioning whether the averaged RDCs correspond to the average of the RDCs calculated for the individual conformations.

Another important issue for the reliability of the calculated averaged data concerns the accuracy of the predicted observables for the individual molecular conformations. Data are usually back-calculated from molecular structures through theoretical models, as in the case of diamagnetic RDCs (through PATI,[34] for steric alignments, or PALES,[35;36] for either steric or electrostatic alignments). The accuracy of these models with respect to the many orienting media that can be used to generate several sets of data, especially in the presence of contributions from electrostatic alignment, is still to be carefully analyzed in the presence of conformational heterogeneity. An alternative “tensor-free” method, called ϑ method, was recently proposed, based on the direct dependence of the RDC between two atomic nuclei on the angle between the internuclear vector and the external magnetic field.[37]

SAXS profiles have been largely used to assess the presence of molecular conformational variability, as they depend on the overall molecular shape. They can be nicely back-calculated through programs like CRYSOL[38]. Although these data do not contain structural details on the nuclear positions, they are very informative on the overall changes of molecular shapes and on the range of the radii of gyration that the molecule must sample. SANS is also sensitive to the overall molecular shape, with the variant that, upon the use of deuteration of domains/subunits and changes in the D_2O/H_2O ratio of the solvent, the scattering contribution of individual domains/subunits can be tuned, and even completely erased [39].

For observables like PREs, relaxation measurements and FRET, further considerations should be made. All these observables arise from modulations of dipolar interactions, and are measured from a signal decay over time. If the rate constant of this decay is larger than the rate of interconversion between different molecular conformations, the observed overall decay is the sum of different decays and therefore it is not monoexponential, and an average value is not available. Otherwise, if the rate constant is smaller, the averaged observable is given by the average of the values corresponding to all sampled conformations, as for the cases described above. Differently to the previous cases, however, the values to be averaged, related to the different conformations, do not depend only on the molecular structure but also on motional parameters, like correlation times and order parameters.[25;40-43] The correlation time which is needed for the prediction of PREs is in fact the shortest between the electron relaxation time, the molecular tumbling time and the time of interconversion between different conformations. If the overall reorientation time of the molecule is longer than, or of the same order of, the rate of interconversion between different conformations, then the interconversion time enters the definition of the correlation time. In such a case, it is not possible to assign a back-calculated value to a single conformation independently of the whole dynamics of the molecule. Tools have been developed for the deconvolution of contributions from overall tumbling and from local dynamics for the analysis of heteronuclear relaxation data and FRET, which is not an easy task especially if the rotational diffusion is not isotropic.[44;44;45]

For PRE measurements induced by the presence of attached tags or spin labels, it is important that these moieties are conformationally rigid.[46] Understanding spin label conformational variability is also important for the analysis of DEER measurements[47] (see also references [48;49] in the present issue).

In the case of two domain proteins, the correlation times for PRE calculations can be the overall tumbling time, the reorientation time of one domain with respect to the other, and the correlation time for the local mobility, in the assumption that the electron relaxation time is longer (as, for instance, in the case of spin labels and some slow relaxing paramagnetic metal ions). Because a description of the dynamics of the molecule is still a forefront field of research, such information is usually missing, and model-free approaches are usually applied. They require the definition of the correlation times related to the different kinds of motion and of the corresponding order parameters, which are quantities often difficult to estimate. Furthermore, all dynamics modes (global, interdomain, local) must be statistically independent from one another: while this seems usually the case for global and local mobility (occurring on time scales differing by several orders of magnitude), interdomain mobility, which can occur on time scales longer, shorter or of the same order of magnitude of the global reorientation time, is often hard to show to be uncorrelated from molecular tumbling.[19;42]

In this issue, NMR-based paramagnetic relaxation interferences (PRI) are proposed for the observation of concerted motions in intrinsically disordered proteins.[50] These restraints arise in doubly spin-labeled proteins and provide information on the proximity of the two spin labels due to the occurrence of dipole–dipole cross-correlation/interference effects.

In multidomain systems, heteronuclear relaxation data measured for the different domains can be analyzed to determine the relative rotational diffusion tensors: comparison of the latter can indicate the presence of, and provide information on, the experienced conformational heterogeneity.[51-53] This approach worked well for instance in the case of the protein diubiquitin, when domain reorientations do not significantly alter the overall shape of the molecule and thus its rotational diffusion tensor, and conformational exchange is limited to interconversion between two states, in the presence of statistically independent overall tumbling and interdomain mobility.[25] More commonly, relaxation data, containing information on the residue-by-residue local mobility, are used to obtain information on the amplitude and frequencies of structural fluctuations occurring on time scales faster than the molecular tumbling time.[3;54]

Differently from the other restraints, PCSs and pRDCs of individual structures can be backcalculated from the magnetic susceptibility tensors which are derived from the experimental data measured for a domain which is known to move rigidly together with the metal ion. Therefore, these data can be obtained independently from any theoretical model, except for the assumption of lack of large internal mobility for the domain containing the paramagnetic metal.[55] Therefore, in order to predict pRDC and PCS values for a molecule composed of multiple rigid domains, the magnetic susceptibility anisotropy tensors (and metal ion's position) can be first determined from the experimental data collected for the domain to which the paramagnetic metal is rigidly bound, and then used to predict the values for the other domains.

Detecting the presence of conformational heterogeneity for a molecule composed of multiple rigid domains from pRDC and PCS is thus straightforward. In fact, in the presence of mobility, the magnetic susceptibility tensors obtained from the analysis of PCSs and pRDCs do not agree for the different domains. Diamagnetic RDCs can also indicate the presence of conformational variability from the disagreement of the tensors corresponding to the individual domains,[23] although to a different extent.[21] PREs can provide precious information because in the presence of conformational heterogeneity the values measured for some residues can be much different than expected for any single conformation. The lack of significant intermolecular PREs and PCSs may also provide indication of the occurrence of extensive averaging among very different conformations.[56;57]

Application of the “tensor-free” ϑ method can permit the simultaneous determination of intradomain and interdomain dynamics in multidomain proteins with flexible linkers, as well as the characterization of the mobility of highly flexible proteins not composed of large rigid segments. The method can in fact describe the experimental RDC values as the average of RDCs of multiple structures, without the necessity of defining tensors by assuming domain rigidity.[37;58] The method can be extended to the analysis of PCSs, although the latter do not arise from partial alignment, but rather from the dipolar interaction between the nucleus and the average magnetic moment of the electron, which is anisotropic if the magnetic susceptibility of the molecule is anisotropic. As a caveat, when a single trajectory is considered only axial orientation can be faithfully reproduced; rhombicity can only be accounted for by the inclusion of several replicas.

Parameterization of the space of allowable states

As mentioned above, every observable which can be used to reconstruct the solution involves temporal and/or spatial averages. As a consequence, individual molecular trajectories are not recoverable. A solution can thus only be thought of as a probability distribution $p(x)$, where the vector x parameterizes the set of allowable states X.

This set can be parameterized in different ways, depending on the unknowns chosen for describing the conformation of the molecule (Fig. 3). In principle, these unknowns should be the Cartesian coordinates of all atoms of the molecule. This choice is however definitely unpractical, since most of these variables are either not related to the available measurements, or strongly correlated with one another (if not completely determined) by some constraints. Hence the aim is to reduce the number of unknowns to the minimum.

The set of variables can be, for instance, the Cartesian coordinates of the nuclei for which experimental data are available, the dihedral angles of the protein backbone, or the rototranslational parameters determining the spatial position and orientation of rigid protein domains. The smaller the number of parameters, the easier the numerical treatment of the data. However, choosing a low number of parameters automatically limits the conformational variability of the molecule to a subset of the states which could actually be sampled. This generates an approximation (inherent in the model) which should be taken into account to decide about its acceptability.

Since the range of variability of each parameter x has bounds (due to spatial restrictions and to the periodic conditions, for angular parameters), any minimization used to find the solution has a (not necessarily unique) absolute minimum in the set of allowable states. Of course, this does not guarantee that any numerical procedure converges to the absolute minimum, because experimental multivariate functions are normally fuzzy, and present many local minima whose value is hardly distinguishable from the absolute minimum.

Each observable $m_i(x)$ is a function of the state x of the molecule. The averaged value $\langle m_i \rangle$ can be represented by the integral over X of the observable against the unknown probability distribution $p(x)$

$$\langle m_i \rangle = \int_X m_i(x) p(x) dx \qquad (3)$$

where dx stands for the appropriate measure for the set X.

Note that averages tend to reduce the variability of the observables. This property may be used to estimate the degree of conformational variability of the molecule.

If all allowable states x were equally probable, then each averaged observable would be

$$\overline{m}_i = \frac{1}{\int_X dx} \int_X m_i(x) dx. \qquad (4)$$

By comparing the distribution of the $\{\overline{m}_i\}$ versus the experimental values $\{\langle m_i \rangle\}$, the degree of asymmetry with respect to the uniform distribution can be estimated (Fig. 4).

Using the set X we can define the set V of convex combinations of observables corresponding to the different states of X

$$V = \left\{ m_i = \sum_{j=1}^{n} p_j m(x_j) : x_j \in X, p_j \geq 0, \sum_{j=1}^{n} p_j = 1 \right\} \qquad (5)$$

The set V is normally a convex set. This is because if two states are allowable, then also the weighted average of the corresponding observables is allowable. Note that these averaged observables may, in some cases, also be obtained from a single "average" state (even if the latter can be unphysical). The Carathéodory's theorem states that any point of a compact convex set can be reconstructed by a convex combination of a finite number of extremal points of the set. The extremal points are a subset of the observables taken when the molecule is in the state $x \in X$. This guarantees that the observables calculated by integrating any probability distribution on the set of allowable states can be obtained by averaging the observables of a finite number of states, even if the unknown probability distribution were thought as a continuous function. With mild hypotheses, we can also drop the weights from the convex combination and suppose that the observables are obtained as the arithmetic mean of the observables calculated for the single structures (i.e. each structure is equally weighted). In other words we are sure that, for any probability distribution $p(x)$, there exists a finite set of structures parameterized by vectors x_j such that

$$\langle m_i \rangle = \int_X m_i(x) p(x) dx = \frac{1}{n} \sum_{j=1}^{n} m_i(x_j) \qquad (6)$$

for all i.

Using the above equality we can determine the ensemble of structures parameterized by the vectors x_j which matches the experimental data, because recovering the continuous probability distribution $p(x)$ would be a hard task even in the simplest cases.

To summarize, in the presence of averaged observables, one may a) estimate the extent of the averaging experienced by the molecule by comparing the range of the experimental observables against i) that obtained by back-calculation over a single conformation, because averaging reduces variability (Eq. 3) and ii) that obtained by back-calculated data over all possible conformations (Eq. 4); b) spot the regions where asymmetry can be located, by use of a finite number of conformers to fit the experimental averaged data (Eq. 6).

Again we stress that, dealing with an underdetermined problem, we are sure that there is at least one solution able to reconstruct the measurements, and indeed there are infinite solutions if no further constraints are imposed to the ensemble. The task of all developed methods is determining how to choose a solution and to assess its properties.

Degrees of freedom, under-restraining and over-restraining

When dealing with an underdetermined problem one of the most interesting and delicate topics is the balance between degrees of freedom of the solution and accuracy of the solution. The number of degrees of freedom of the solution, i.e. the number of unknowns needed to define the ensemble which will be used to simulate the observables, is normally a multiple of the number of conformations present in the ensemble. Of course, the agreement with the data tends to improve (or anyway to stabilize) by increasing the number of conformations. In a linear approximation, when the number of degrees of freedom exceeds the number of independent measurements, a solution able to reconstruct exactly all consistent experimental data can be always found.

Since the experimental data are normally affected by errors, reconstructing exactly the noisy measurements means reconstructing the errors as well, which is something we would like to avoid. This raises the question of how to balance between degrees of freedom of the solution and accuracy of the solution. The best agreement between the experimental data and the averaged observables calculated from the structural ensemble can be obtained by minimizing a target function (TF), typically defined as a weighted least-square difference between experimental and back-calculated data. This is a non-negative function which goes to zero when a perfect agreement is achieved.

The first point, which is common to most of the methods used to determine a solution, is to decide whether a certain distribution can be considered as a solution. Since the data are affected by errors, a solution should be defined as a convex combination of structures such that the TF will not exceed a threshold. This threshold may be based on the absolute minimum of the TF, obtained when no restraints are imposed to the structures, or can be determined from the estimated experimental errors via some statistical tests (see later). The threshold level can of course influence the final distribution, irrespective of the applied method.

The standard technique to face the problem of balance between degrees of freedom of the solution and accuracy of the solution, when dealing with ill-posed problems, is the Tikhonov regularization. Loosely speaking, there will be a second function quantifying the undesired properties of the solution, which is called a penalization term (PT). The penalization term should be as low as possible (see Fig. 5). Most of methods used in ensemble averaging protocols fit into this frame, with the PT appearing as an energy term, meaning that we are looking for low energy solutions. The Tikhonov regularization finds the solution as the minimizer of a linear combination of the TF and the PT.

To facilitate the understanding, the example of recovering the MEP solution is here discussed:

i) the TF must reflect the agreement between the model ($f_j(x_i)$ where x_i indicates the i conformation) and the experimental observables (m_j with standard deviation σ_j), for instance $TF \propto \sum_{ij} \left(f_j(x_i) - m_j\right)^2 / \sigma_j^2$;

ii) we want to impose adherence to a force field, which could provide an approximate estimate of the energy of each conformation ($\tilde{E}(x_i)$).

Using a discrete approach providing ensembles of N equiprobable states x_i, the constrained energy term $\sum_i \tilde{E}(x_i) + \frac{kT}{N} \sum_j \left(f_j(x_i) - m_j\right)^2 / \sigma_j^2$ is then minimized. It can be demonstrated[59;60] that if a large number of conformations N is used, then the solution corresponds to the maximum entropy solution, because the constrained energy gives rise to a Boltzmann distribution $\tilde{Q}_N(x) = \frac{1}{Z_N} e^{-\frac{1}{N}\Sigma_i \tilde{E}(x_i)/kT} \prod_j \delta\left(\frac{1}{N}\Sigma_i f_j(x_i) - m_j\right)$, where Z_N is the partition function which normalizes the integral of the distribution to 1. In turn, this corresponds to the minimizer of the Kullback-Leibler divergence (or relative entropy) $h_r(P) = \int_{x\in X} P(x)\log\left(\frac{P(x)}{\tilde{Q}(x)}\right) dx$, where $P(x)$ is the probability distribution.

In most LW methods the PT is implemented by penalizing the number of structures of the ensemble if different from a preset value, in order to obtain solutions composed of conformers with large probabilities. The W_{max} for each conformer is obtained by penalizing the sum of the probabilities of all other structures comprised in the ensemble, in order to maximize the probability of the structure for which we want to determine the W_{max} value.

The aim is to reconstruct the data as well as possible (i.e. minimize the TF) with a solution which minimizes the unwanted features (i.e. minimize the PT). The two goals are normally in contrast, so a balance between the two should be decided. If the minimization of the TF is privileged, the solutions fit the data very well, but they contain features which do not reflect only the physics of the system but also the random features of the experimental error. On the other hand if the minimization of the PT is privileged, the solutions possess the desired properties but they may not fit the data nicely.

Increasing the number of conformations considered in the ensemble not only usually increases the agreement with the data, but also minimizes the PT, unless the latter contains a more or less explicit dependence on the number of conformations. Then, in principle, the best results are obtained when the number of conformations is large (as for instance in solutions of the MEP class). The large number of structures to be considered is however a problem for the numeric minimization, because of the possibility of finding local minima of the TF instead of the global minimum which is required. Moreover, unless the entropy is controlled in some way via the PT, the calculated structural ensembles may have features which are present in specific solutions and absent in other solutions, thus implying that they cannot be inferred from the measurements. This is typical of the so called under-restrained problems. On the other hand, if the number of structures considered in the ensemble is not sufficient, the agreement with the measurements will not be optimal, and over-restrained solutions will be obtained[61;62]. As a result, methods of the MEP class tend to suffer from under-restraining, while methods involving LW solutions tend to suffer from over-restraining.

Of course, this implies that the most efficient way of avoiding over-restraining are the MEP approaches. A solution with a large entropy contains by definition a large uncertainty, hence every feature shown by the MEP solution is relevant. On the other hand, as already mentioned, the large number of structures to be considered implies that one has to tackle very difficult computational issues.

Protocols to avoid both over-restraining and under-restraining are based on gradually increasing the number of structures included in the ensemble (see Fig. 6), and devising some statistical test in order to decide when to stop. In these approaches, the PT constrains the number of conformations to be considered in the ensembles. One may stop for instance when the number of conformations is the smallest compatible with the data, i.e. the solution has a TF not exceeding the defined threshold, or when the addition of a new structure does not change the TF. The threshold level can be checked with the χ^2 statistics, if the variances and covariances of the measurements can be estimated. The expected χ^2 and its standard deviation depend on the number of degrees of freedom of the ensemble. A Student t-test can decide within a certain confidence level if the addition of a new structure affects the TF significantly, although in practice the uncertainties, including the modelling errors, are often difficult to be estimated.

Alternatively, cross-validation can be used. In this approach a fraction of the data (the free data) is excluded from the TF and the agreement between the measured and back-calculated data of the excluded fraction is checked. In the case of over-restraining (the number of conformations is too low), if the number of structures in the ensemble increases, the agreement on the portion of the data defining the TF generally improves. However, since the free data are excluded from the TF, when the ensemble is too large, the freedom of the free fraction of the data increases so much that the agreement becomes worse, and thus under-restraining can be detected.

Finally, the entropy of the system can be calculated. As described in Choy et al.,[63] an artificial entropy term can be added to the total energy, so that population weights distribute among similar conformers instead of concentrating on individual conformers. This increases the number of conformers in the final ensemble that have significant population weights. If the entropy does not increase when new structures are added to the ensemble, we can deduce that we have reached the maximum entropy of the system.[63-66]

While based on different statistical quantifiers, the approaches for checking under-restraining can be resumed by Occam's razor[15] (something which is not needed should be cut out), or by the popular motto "the lesser the better" which is the cardinal idea when dealing with ill-posed problems.

Is there a theoretical limit to the maximum number of conformations to be included into the ensembles in LW approaches? The maximum limit of the ensemble size should be constrained to no more than the effective rank of the prediction matrix A of the predefined pool of structures[21;67]. The matrix A is defined so that each column contains the values back-calculated for a protein conformation corresponding to the experimental observables; the different columns report the predicted data for all different possible protein conformations. The rank of A results as the number of "large" (e.g., greater than 0.01[21;67]) relative singular values, $\sigma_i / \sigma_{\mathrm{max}}$: trying to recover a larger number of conformations would result in overfitting and would introduce conformations that are not really needed to reproduce the data.

A summary of the main ensemble averaging protocols

As already mentioned, all proposed protocols provide solutions as probability distributions of the allowable states, but to overcome the problems of approximating a continuous probability distribution, the solutions are found as the sum of a finite number of single, often equiprobable, states. Thus, ensemble averaging is performed in all the methods commonly implemented. The techniques proposed are however very different also depending on the predominantly applied MEP or LW approaches.

Figure 7 summarizes the main different protocols described below in detail, depending on the method used in generating the conformations belonging to the retrieved ensembles (full atomistic determination of nuclear coordinates, such as molecular dynamics, stepwise sampling, free domain rototranslations, see Fig. 3) and on the predominantly applied MEP or LW approaches.

Replica averaging minimization/Restrained molecular dynamics

Structure determination is usually achieved by minimizing a hybrid target function which contains both the agreement to a physical force field and to the experimental data. In the presence of mobility, the experimental data should not match the values back-calculated for a single conformation, but rather the averaged values from an ensemble of conformations. Therefore, ensembles of structures can be determined by simultaneously searching for multiple replicas providing agreement with the averaged experimental data. This method introduces a new parameter in the structure calculation protocol, i.e. the number of parallel replicas to be used to reproduce the experimental data. As discussed previously, in LW approaches this number is usually set using the Occam’s razor principle, implying that the number of replicas to be considered in the ensemble should be the lowest possible needed for a satisfactory agreement with the experimental data (Fig. 6).

The most representative conformations of the system are thus supposed to be determined by constructing ensembles through the following protocol[15]:

1. An ensemble is built with a fixed number of structures (usually with the same weights) with conformations chosen for a best agreement with the experimental data (lowest TF);
2. the number of structures is changed, and a new ensemble is built in best agreement with the experimental data, i.e. with its own lowest TF;
3. the most informative ensemble is then determined as the one with the lowest number of structures, providing an agreement with the experimental data not significantly worse (i.e. with TF not significantly larger) than that with one additional structure. As already discussed, the criterion used to determine the optimal number of structures can be determined by χ^2, student t-test, L-curve statistics, or by a limiting threshold calculated with respect to the lowest TF determined with a relatively large number of structures.

This procedure ensures that the selected ensemble is in best agreement with the experimental data and prevents from inclusion of structures which are not required for the fit of the data. Of course, this does not mean that the data cannot also be fit with a larger number of conformations which can be even more different among them. Rather, the Occam's razor suggests that the current data provide no basis for invoking a larger conformational ensemble.[15] Alternatively, as already discussed, the appropriate value of replicas to be used in the ensemble can be determined by cross-validating with independent data not used in the structure determination.[60] Replica averaging minimization methods can be very demanding from a computational point of view due to the many local minima that may prevent the minimization to reach the global minimum, unless the whole conformational space is carefully mapped.

A method used to generate best-fit ensembles is complementing experimental data with a priori information derived from molecular dynamics.[61] Molecular dynamics provides the free energy landscape, which is expected to correlate with the statistical weights of various conformations. In restrained molecular dynamics approaches, the energy function is perturbed with a term driving protein conformations towards structural models in agreement with the experimental data. Multiple replicas of the protein must be simulated in parallel at each point in time and a restraint in the energy function is added as a function of the agreement between experimental data and back-calculated averages from all of the replicas.[3;68] Annealing cycles between e.g. 300 and 400 K are included in the construction of the ensembles[3;59]. Again, the number of replicas to be considered for averaging should be carefully considered because if too large the experimental information is insufficient to define the structure of all of the replicas (under-restrained solution).

Alternatively, replica averaging can be performed using MEP approaches, which consist in selecting, among the infinite number of distributions compatible with the data, that providing the largest degree of uncertainty of the variables of interest.[69;70] The Kullback-Leibler divergence is used in place of the MEP when conformational variability occurs through fluctuations around one average conformation, it therefore yields good results for well-folded proteins when conformational fluctuations are modest.[71] The maximum entropy approach provides a way to ensure agreement with the experimental data without adding further information on the conformational distribution that is not carried by the experimental data.

It was shown that replica-based calculations converge to the maximum entropy solution[59;72] when the number of replicas goes to infinity and their weights are appropriately constrained. Therefore, the MEP solution can be determined as the limit case of a number of replicas approaching infinity. Incorporation of experimental data as replica-averaged structural restraints in molecular dynamics simulations, given an approximate force field, can thus provide an accurate representation of the Boltzmann distribution of a system.[59;72;73]

Of course, the true MEP solution is in general a continuous probability distribution. The maximum entropy approach can thus be implemented in a way to obviate the need to simulate many coupled replicas.[73] Since the use of large numbers of replicas may prove to be computationally intractable or impossible, approaches which are independent of their number are developed.[71]

Sample-and-select (SAS)

Determining protein ensembles through multiple replicas minimization in restrained molecular dynamics calculations can be very expensive from a computational point of view. The SAS approach was thus developed, based on selecting a number of conformations from a predetermined pool of structures, in order to construct ensembles providing back-calculated averaged data in agreement with the experimental data[65]. The predefined pool may have been calculated from statistical models taking into accounts the backbone dihedral angle variability (as in Flexible-Meccano[74] or Ranch[75]), from molecular molecular dynamics simulations without any inclusion of the experimental data in the energy function, often using techniques enhancing the conformational sampling (high temperature molecular dynamics or simulated annealing, accelerated molecular dynamics,…), or from a stepwise geometric sampling when rigid domains can be defined in the molecule (see next section). The success of the approaches based on selecting conformations from a predefined pool depends on the completeness of the pool both in terms of structural variability and resolution.

The programs developed to select subsets of conformations (either with the same weight or with different weights) providing ensembles in agreement with the experimental data are numerous (ASTEROIDS,[76] ENSEMBLE,[63] SAS,[77] EOM,[75] MES,[78] EROS,[64] SES,[67]…). Among them, some programs (ENSEMBLE, EROS) implement the maximum entropy weight distribution to determine representative ensembles.[63;64]

In the SAS approach, it is commonly assumed that pooling together ensembles of structures, each of them in agreement with the experimental data, can provide a statistical description of the major conformers of the real ensemble. As already discussed, the fact that an ensemble is in agreement with the experimental data does not guarantee that it is accurate. It was shown through numerical simulations that it is possible to determine best-fit ensembles from synthetic data which do not contain the conformations used to generate the synthetic data (or similar conformations).[66;79] This suggests that great care should be taken when interpreting the best-fit ensembles, as the conclusions may be significantly biased by the employed numerical methods and the analyzed structural features. Common approaches actually consist in constructing multiple best-fit ensembles and looking for common characteristics[80;81] to identify recurring structural features:

1. Conformations are selected from a predefined pool through a simulated annealing or a genetic algorithm search to construct ensembles, with the smallest reduced χ^2 statistics.
2. The number of conformers present in the ensembles is usually, but not always, determined as the smallest number needed to obtain a good agreement (e.g. the minimum χ^2) with the measured data. Choosing a very large number of structures would imply passing from a LW-based approach to a MEP-based approach, if the entropy is simultaneously maximized.
3. Many ensembles can be produced by repeating this procedure hundreds of times.
4. A final ensemble is finally built by pooling together all the ensembles of conformations selected over all runs. In this ensemble, conformations occurring in several of the starting ensembles will obviously have a larger weight.
5. Therefore, from this final ensemble the statistical weights of dominant conformers is captured.

Only the overlapping fraction of conformations selected in many cycles are expected to be relevant. This approach would give confidence that the preserved structural features (like radius of gyration, distance maps, etc.) are accurate. However, it is difficult to determine how many ensembles should be analyzed and to which percentage the preserved structural features should be identified, to exclude that these features are only accidental. The entropy of the final ensemble can also be calculated. The ratio of the entropy of the final ensemble versus the entropy of the whole pool is a measure of the effectiveness of the experimental restraints.[65]

In a slightly different approach, unbiased molecular dynamics simulations (without any inclusion of the experimental data in the energy function) can be performed and the calculated conformations can then be reweighted using the maximum entropy principle, in order to determine ensembles in agreement with the experimental data.[82-85] In the LW approach, the population of a limited number of X-ray structures, possibly complemented by few structures calculated with molecular dynamics simulations, can be determined to fulfill the experimental restraints, as performed in this issue for the characterization of HIV-1 protease conformational sampling with and without inhibitors.[48]

When SAS approaches are applied to the study of intrinsically disordered proteins, ensembles from few tens to few hundreds conformations can be obtained and validated to reproduce some experimental parameters, like the radius of gyration, or by cross-validation of data not employed in the fit.[75;76;86] The presence of minor populations of conformers can also be enlightened by building contact maps determined on a large set of LW ensembles.[87]

The SAS approach can be conveniently used also for modeling the spatial distribution of spin labels, which can be highly flexible. In this issue it is shown that determining the most favourable conformer orientations using intra-molecular PRE data before performing docking calculations based on intermolecular PREs can improve the accuracy of the results.[46]

Sample-and-select in multidomain systems

In the approaches described above, the nuclear coordinates of all atoms are independent, except for the presence of the covalent bonds. In proteins and nucleic acids composed of domains with a low internal mobility, it may be convenient to avoid considering both the internal variability of protein domains and the interdomain conformational rearrangements. Approaches were thus introduced assuming rigidity of the individual domains and allowing all interdomain rearrangements not resulting in steric clashes. In these cases, a geometrical coarse grain pool of conformations is built by varying in steps the rototranslational parameters of the moving domains; of course, it should be ensured that the resolution of the geometrical pool is high enough. All conformations that are stereochemically impossible to achieve either because the linker is too short to maintain connectivity or because it leads to severe steric constraints should be removed from the pool.

In summary, when the conformations are selected from a predefined pool of structures, the pool should be built either comprising representatives of all sterically allowed conformations determined by free rototranslations of the rigid domains composing the biosystem, or all conformations broadly sampling the inter-domain free energy landscape, or the conformations calculated from long molecular dynamics simulations. Using a geometrical coarse grain pool, of course, it is not possible to consider any intra domain structural variability. As already discussed, the use of molecular dynamics simulations for the generation of the pool introduces a bias in the calculated ensembles, and the quality of the result depends on the reliability of the molecular dynamics. On the other hand, molecular dynamics has the advantage of taking into account internal structural changes of the domains that cannot be considered if a geometric pool is used.

Sparse Ensemble Selection

To avoid overfitting of the data, small-sized ensembles should be constructed. Along these lines, a particular implementation of SAS is the Sparse Ensemble Selection (SES) approach. SES was developed to select the smallest (sparsest) nonuniformly weighted representative ensemble which explains the experimental data from a predefined pool.[67] The search for the conformations is sequential and is based on the following steps:

a) conformations are ranked according to their compliance with the experimental data;

b) the best-scoring subset of the conformations is selected;

c) each conformation selected in the previous step is complemented with all other conformations, one at the time; all pairs are then ranked according to their compliance with the experimental data;

d) the best-scoring subset of the pairs is selected.

Steps c-d are repeated using triplets, quarterts etc. of conformations selected, until the target function does not change significantly upon addition of further conformations, as evaluated using the L-curve method.

In order to maximize the weights of a very small number of conformations, solutions with total weights something less than 1 are accepted. This is equivalent to replace each large weight conformation with combinations of neighboring conformations. An example of the application of the SES approach is provided in this issue by the study of the conformational variability in di-ubiquitin.[88]

MAP, MaxOcc, MaxOR and MinOR in multidomain systems

A different approach was introduced to analyze the molecular conformations with respect to their compliance with the experimental data, i.e. their maximum probability. Maximum Allowed Probability (MAP)[17;18] and Maximum Occurrence (MaxOcc)[89] calculations provide the maximum weight that any conformation can have whatever the real ensemble to which it belongs to (Fig. 8A). The MAP approach uses free domain movement, while the MaxOcc method selects the protein conformations from a predefined pool in order to solve the computational issues in the calculation of the solution when different experimental datasets are used jointly. MAP and MaxOcc indicate how much a given conformation can contribute at maximum to the experimental average.[79;89-92] They thus provide an estimate of the W_{max} value described in the Introduction. Calculations are performed by searching a conformational ensemble which includes the given conformation with a fixed weight and tens of other conformations, selected with a minimization program in order to provide averaged data in agreement with the experimental data. These calculations are repeated for increasing weight of the selected conformation, until it becomes impossible to find an ensemble in good agreement with the experimental data.

In order to maximize the weight of the conformation under analysis, MAP and Maxocc methods thus minimize the weight of the other conformations present in the ensemble (i.e. their weights represent the PT which is minimized). It is clear that in order to obtain the largest weight for one structure, the other structures completing the ensemble should be placed as far as possible from the first one. Therefore, although the MaxOcc result for one conformation is a solution, the obtained ensemble of structures should not be regarded as a reliable conformational ensemble which can be used to represent the system: what the calculation provides is a safe and reliable estimate of the maximum weight of that conformation.

It is often assumed that the conformations with the largest W_{max} are those which candidate as the most representative for the system. Unfortunately, there is not a straight correspondence between MAP or MaxOcc and real weight because LW conformations may include ghost solutions,[17] arising from the degeneracy of the parameters (like pseudocontact shifts and residual dipolar couplings) with respect to the molecular structure, or poorly sampled conformations which are structurally averages of conformations with large weight. In this second case, the molecule switches between structurally very different conformations.

The MAP and MaxOcc approaches can be also extended to the study of ensembles of structures, as those populating regions defined within the conformational space possibly sampled by the system. The aim is to identify the conformational regions with W_{max} equal to 1, i.e. the regions where a full solution can be constrained, or the conformational regions with W_{max} close to 0, i.e. the regions where no structures can be placed with a significant probability.[93] The calculations of the maximum and minimum occurrence for regions, MaxOR and MinOR,[94;95] 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 weight (Fig. 8B). The MaxOR values can then be calculated over *pairs* of regions, in order to detect bimodal distributions. Again, the selected regions with high MaxOR, as in all LW approaches, address only one of the possible solutions, i.e. the simplest solution describing the asymmetries present in the experimental data.

In this issue, it is shown that the MaxOcc and MaxOR analyses performed for HIV-1 TAR RNA, consisting of two helical domains connected by a flexible bulge junction, identify the most likely sampled region in the conformational space of the system, which strikingly overlaps well with the structures independently sampled in unbiased molecular dynamics calculations (without any inclusion of the experimental data) and even better with the SAS ensemble.[95]

Detecting minor conformations

Due to the sixth power dependence of PREs on the distance between the nucleus and the paramagnetic center, the paramagnetic broadening affects only residues in close proximity, although transiently, to spin labels. This is why paramagnetic spin labels can detect the presence of low weight conformations together with a known predominant conformation of the system. In fact, if the known predominant structure of the system is not consistent with the experimental paramagnetic broadening, minor conformations with nuclei close to the paramagnetic center, although sampled only for a short time, can be identified. For instance, if a protein can have different conformations, i.e. compact or extended, in different conditions, it is possible to monitor whether the two conformations can coexist in fast exchange with different weights. In this way, it is possible to determine whether a protein experiences an equilibrium between open and closed conformations, even when the weight of the latter is as low as 5%.[96;97]

Analogously, the paramagnetic relaxation enhancements measured for a protein-protein complex can indicate if the complex spends a fraction of the time in an ensemble of conformations different from the crystal model[20;98;99] or from the NMR structure determined with conventional approaches.[100;101] An example is provided in this issue by the case of the complex between cytochrome c and cytochrome c peroxidase.[46]

Geometric interpretation of ensemble averaging approaches

Figure 9 summarizes the abovementioned protocols for ensemble reconstruction, using simple geometrical examples. In this analysis, the datasets are assumed to be internally consistent, i.e. that all the measurements refer to the same conformational distribution of the macromolecules in the sample. Any set of K experimental measurements corresponding to rigid protein conformations defines a space with dimension K, and each conformation can be described as a point in this K-dimensional space. Furthermore, any set of averaged experimental measurements can be seen as a point ($\overline{P}$) which can be reconstructed by a convex combination (i.e. a linear combination with weights summing up to 1) of sets of calculated observables corresponding to a number of different protein conformations. The set of such convex combinations defines a convex set V with dimension $n \leq K$ of the order of the number of non-zero eigenvalues of the experimental data. Referring to panel (a), the set V is represented by the ellipse. Not all conformations are necessarily at the boundary of the convex set (see the grey dots in the Figure 9). Also, not every point of the boundary represents a single conformation. As an example to clarify the dimensions, let us consider the paramagnetic RDCs from a rigid protein domain. Here K is the number of measurements, so a set of measurements represents a point in this space. However it is known that there are only 5 linearly independent RDCs for each paramagnetic tensor, so if we have enough measurements the dimension n of the set V is 5. The RDCs depend only on the orientation of the protein domain with respect to the paramagnetic tensor, so any conformation can be parameterized for instance by 3 Euler angles. Since the boundary of a convex set of dimension 5 has dimension 4, it is clear that the entire pool of conformations does not fill the boundary of V.

If we only consider convex combinations of conformations taken from a large pool of structures we obtain a polyhedron M, shown as the thin dotted lines in panel b. The number of vertices of M is of the order of the number of the possible protein conformations in the pool of structures, shown as dots in the figure. Of course, the larger the pool, the better the approximation of V, in particular the dimension of M should match that of V. Any point $\overline{P}$ inside M can be reconstructed using at most $n+1$ vertices of the polyhedron. A single set of RDCs, for instance, defines a polyhedron with dimension 5. Any point $\overline{P}$ can be reconstructed using at most 5+1 vertices out of all possible vertices of the polyhedron defined by the conformations of the pool.

Protocols to build the conformational ensembles are implemented either allowing the different conformations to have different weights or fixing their weights to be constant; the solution is in any case the mean of the measurements of the conformers.

For the first category, in the example of the figure, the point $\overline{P}$ can be reconstructed using a linear combination of three points P_1, P_2 and P_3. The weight of each P_i is equal to the ratio of the areas of triangles T_i and T , where T is the triangle $P_1P_2P_3$ and T_i is the triangle where the point P_i is replaced by $\overline{P}$ (panel c).

For the second category, the reconstruction can only be the barycentre of the conformations of the ensemble, see panel d. It is known that any point inside V can be reconstructed in this way if all the possible conformations are considered for the combination.

The true MEP solution would be an integral mean of all the conformations, but we can think of the MEP solution as approximated by the replica ensemble method with a large number of conformations. Loosely speaking the entropy can be represented by the area of the polygon in the n-dimensional space defined by the conformations considered. Hence the replicas will try to maximize this area, still maintaining the property that the point $\overline{P}$ is the barycenter of the polygon. As a consequence, one may expect that the density of conformations is thicker near the point $\overline{P}$ and thinner far from it. This property becomes more apparent the farther the point $\overline{P}$ is from the barycentre of the set V (panel e).

The SAS method uses the simulated annealing or any similar technique to determine the ensemble which is more consistent with the data choosing them from a large pre-determined pool. First one needs to define the number of conformers in the ensemble (6 and 4, respectively in panel f), then the calculation is repeated for a large number of ensembles, and the common features of the ensembles are analyzed. Sometimes, the more represented conformations are shown (panel f1).

The Occam's razor approach to determine the minimum number of conformations is illustrated in panel g. Using a single conformation the point $\overline{P}^{(1)}$ is obtained. Using two conformations the point $\overline{P}^{(2)}$ is obtained as the average of the two vertices. Using three conformations the barycenter $\overline{P}^{(3)}$ of the triangle is obtained. A statistical test would determine if the better agreement obtained because $\overline{P}\overline{P}^{(3)}$ is smaller than $\overline{P}\overline{P}^{(2)}$ is worth the increased number of conformations considered.

The SES method consists in reconstructing $\overline{P}$ using a smaller number of vertices, 2 in our example (panel h) that can still yield a solution. While it is not possible to reconstruct exactly the point $\overline{P}$

using less than *n*+1 vertices, we can obtain a good approximation by letting the sum of the probabilities be less than 1: the combination of $\bar{P}_2$ and $\bar{P}_3$ reconstructing $\bar{P}$ as shown in panel b may be seen as a combination of $\bar{P}_2$ and $\bar{P}_3$ with weights scaled down according to the linear scaling factor between the triangles $O\bar{P}_2\bar{P}_3$ and OP_2P_3 (*O* is the origin). In this example, it is easy to see that this ratio is close to 1 when using P_2 and P_3, while it would be smaller using any two other vertices.

The MaxOcc method (panel i) consists in finding, for any vertex P_i, the point Q on the boundary of the polyhedron M, that originates from the continuation of the segment $P_i\bar{P}$, and defining the maximal occurrence of the conformation P_i as the ratio $\frac{\bar{P}Q}{P_iQ}$. In this example, it is easy to see that the MaxOcc is large for P_2 and P_3, and small for P_1. In the case of the MAP method, the crossing point Q' is the intersection with the boundary of the convex set V, so that the maximum allowable probability is defined as $\frac{\bar{P}Q'}{P_iQ'}$.

Conclusions

We have discussed the main methods used for ensemble reconstruction. Loosely speaking, they can be framed into two extremal approaches: LW and MEP. It is usually assumed that the calculated ensembles can provide an estimate of the extent of the conformational heterogeneity. This is not strictly true in general using LW approaches, especially in the presence of an extensive conformational heterogeneity of the system. It may be correct when the mobility is low, if the smallest number of conformations that is sufficient to achieve a good agreement with the experimental data is used. The use of a redundant number of conformations can in fact introduce noise through the inclusion of counterbalanced conformations. On the other hand, the fact that all experimental restraints can be satisfied by one ensemble does not mean that such an ensemble is a unique and complete description of the system; it only describes the information content of the experimental data by enhancing the specific features that they contain.[87] The calculation of W_{max} (as performed through MAP or MaxOcc) then provides an upper limit to the occupation of points or regions of the conformation space.

On the other side, the MEP solution, which can be recovered with a number of conformations approaching infinity, blurs any detailed description of each conformation included in the ensemble, but focuses on determining the minimal set of features common to all solutions.

In conclusion, the two classes of approaches provide complementary views of the system that can both shed light on its conformational variability. Only the availability of information from non-averaged observables or from theoretical tools, such as molecular dynamics calculations, can help determining which of the two approaches is most appropriate to describe the system.

The need for established and well validated protocols for a reliable characterization of the conformational variability of biological molecules is expected to become more and more urgent in the next future, now that its importance for the molecular function has been evidenced. The impact of the studies performed to determine the biological activity of macromolecules by describing their conformational variability is continuously increasing, and mobility studies are now expected to always complement structural studies. Theoretical advancements in this field are continuously developed and a variety of software tools are proposed. These tools should ideally be able to describe the upper and lower limits in the probability distribution of the conformations which can be sampled by the system. The recent developments of molecular dynamics calculations, able to sample fast and carefully the whole molecular conformational space, are expected to provide an essential contribution to this aim.

Acknowledgements

This work has been supported by Ente Cassa di Risparmio di Firenze, MIUR PRIN 2012SK7ASN, European Commission projects pNMR No. 317127, BioMedBridges No. 284209, and the EU ESFRI Infrastructure Instruct through its Core Centre CERM, Italy.

Table 1. Long-range observables for the investigation of conformational heterogeneity

Observable	**Chemical requirements**	**Required for analysis**	**Provides information on**
PRE	Spin label or paramagnetic metal, isotope labeling	Correlation times	Paramagnetic center-nucleus distance
PCS	Paramagnetic metal, isotope labeling	$\Delta\chi$ tensor if not determined from nuclei moving rigidly with the metal ion[a]	Nuclear positions in the frame of the $\Delta\chi$ tensor
pRDC	Paramagnetic metal, isotope labeling	$\Delta\chi$-tensor as determined from PCS/pRDC of nuclei moving rigidly with the metal ion[a]	Nuclear pair orientations in the frame of the $\Delta\chi$ tensor
External alignment RDC	Orienting medium, isotope labeling	Alignment model[a]	Nuclear pair orientations in the frame of the alignment tensor, Molecular shape, local mobility
Relaxation measurements	Isotope labeling	Correlation times, (diffusion model)	Molecular shape, local mobility
FRET	Fluorescent labels	Correlation times	Fluorescent donor-acceptor distance
DEER	Spin labels		Paramagnetic center-paramagnetic center distance
SAXS	Several samples at different concentrations		Molecular shape

SANS	Isotope labeling, several samples at different concentrations		Molecular domain shape

[a]except using the theta method (see text)

Figure 1. Is it possible to retrieve the values obtained by rolling a die a number of times from the averaged value? A probability distributions (for averaged values from 1.5 to 5.5 in steps of 0.5) can be calculated through the MEP approach; the W_{max} values provide the upper bounds for the real distribution, which is correct whatever is the number of times that the die is rolled.

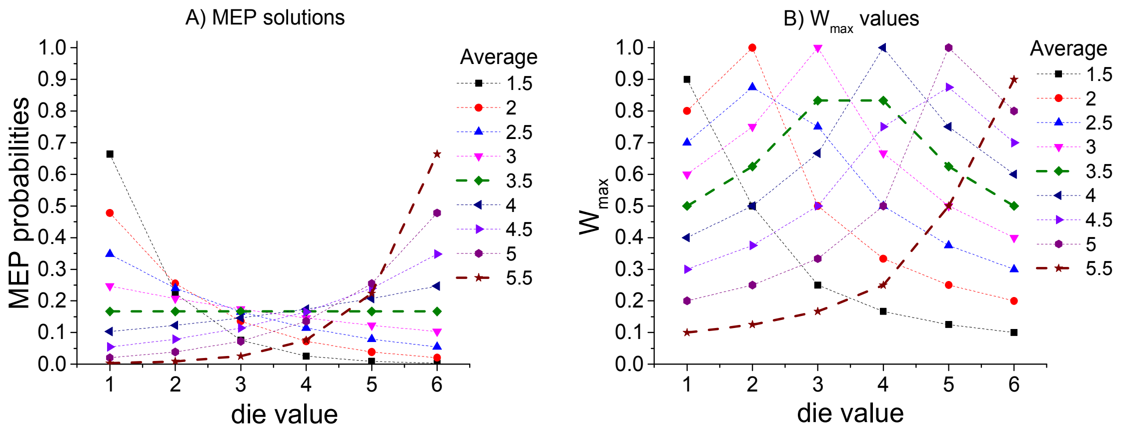


Figure 2. Depending on the number of times that the die is rolled, different solutions can be determined, as shown here for the case that the average of the values is 5.5. The occurrences (expressed in percent) of the different values are compared with the MEP probability distribution and the W_{max}. Bars with the same colors indicate the occurrences for the same of all possible combinations of values providing the same average of 5.5 (the number of possible combinations, of course, increases by increasing the number of rolls).

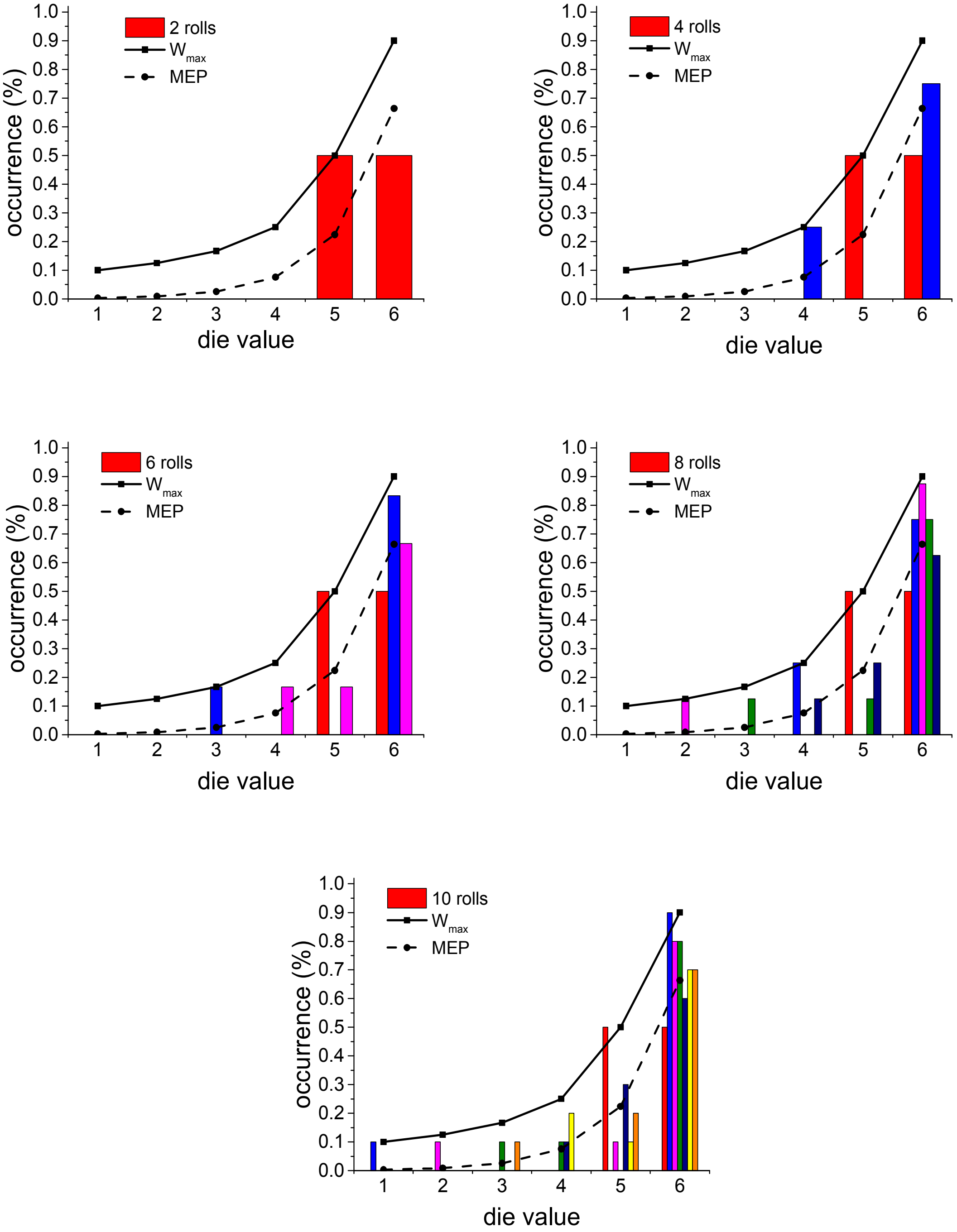


Figure 3. Describing biomolecular conformations, here applied for simplicity to a two-domains protein: a) a fully atomistic description can be applied; otherwise, simplified structural models could be provided by b) parameters such as the dihedral angles of the mobile residues (here representing a linker between the two domains or the c) rototranslational parameters defining the interdomain position.

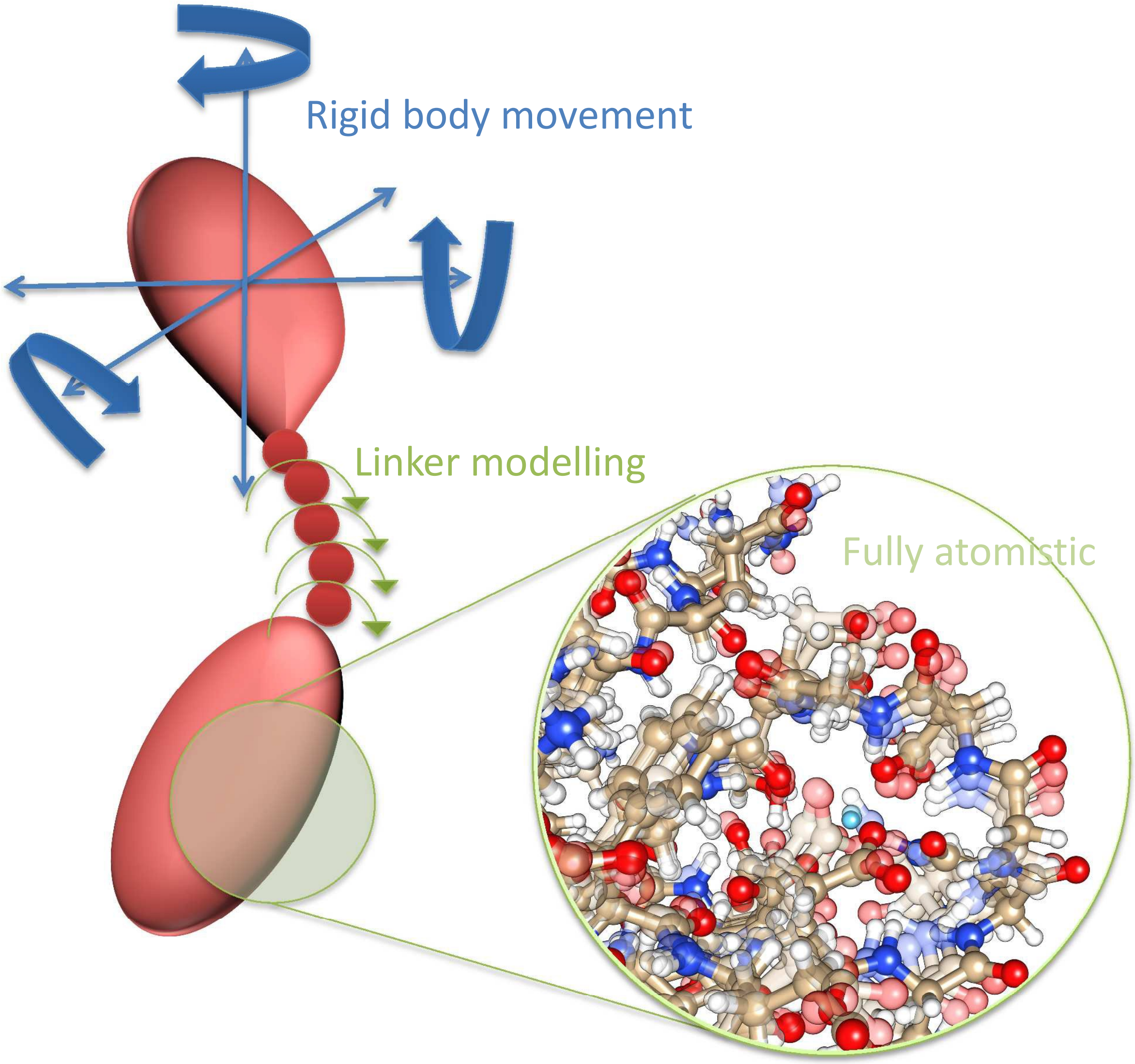


Figure 4. When a paramagnetic metal is introduced in a biomolecule, self-orientation arises. In a two domain protein, assuming the metal is framed in one domain and the RDCs are measured on the other, the following situations can be encountered: (A) a single interdomain orientation is present, (B) an asymmetric distribution of orientations is present or (C) orientations are uniformly sampled. The corresponding distributions of the (averaged) paramagnetic RDCs are shown in panels A, B, or C, respectively.

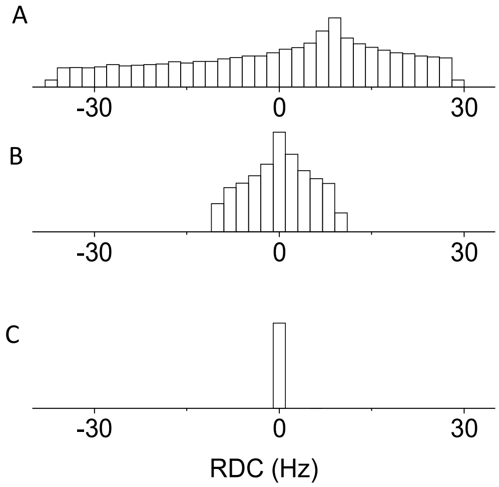


Figure 5. A simultaneous minimization of both TF and PT can be achieved by minimizing the sum of the two, with a weighting factor *k* (TF+*k*PT). The optimal value of *k* can be determined by plotting, in log-log scale, PT and TF for different *k* values. This is the so called L-curve method. In this example, where the PT is given by a sum of the weights of all conformations present in the ensemble different from 1, the optimal *k* is around 10.

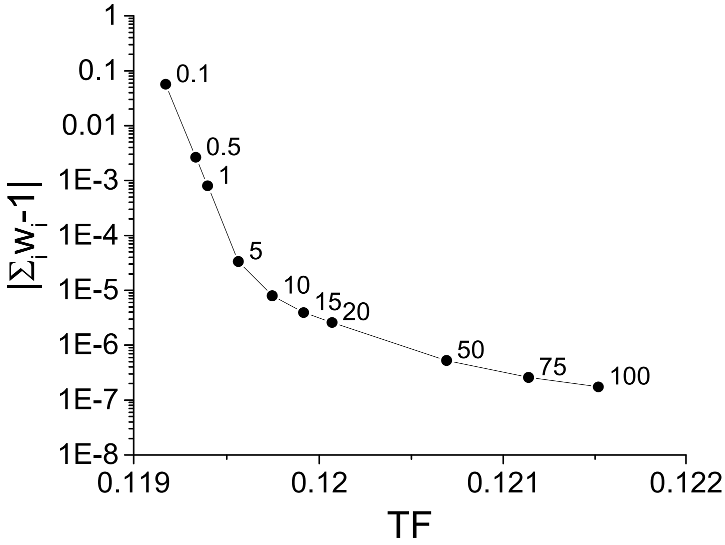


Figure 6. Using the Occam's razor principle, the number of replicas to be considered in the ensemble should be the lowest possible needed for a satisfactorily agreement ($\chi^2_r \approx 1$) with the experimental data (4 in this example).

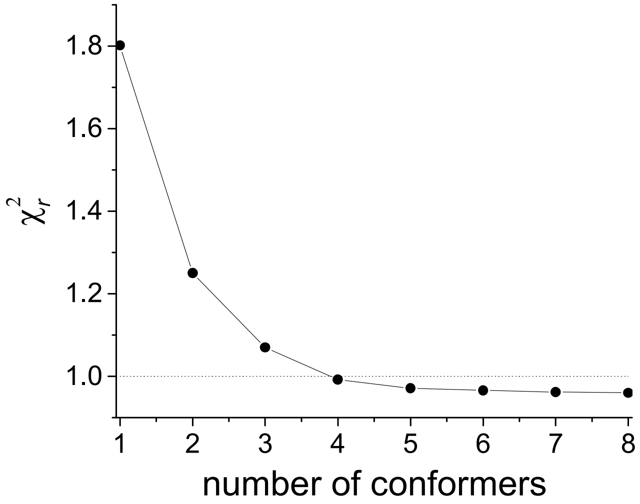


Figure 7. Protocols for analyzing averaged data differ i) in the way the different structures are generated (full atomistic determination of nuclear coordinates, such as molecular dynamics, stepwise sampling, free domain rototranslations, see Fig. 3) and ii) in the approach (LW or MEP) used for building the ensemble.

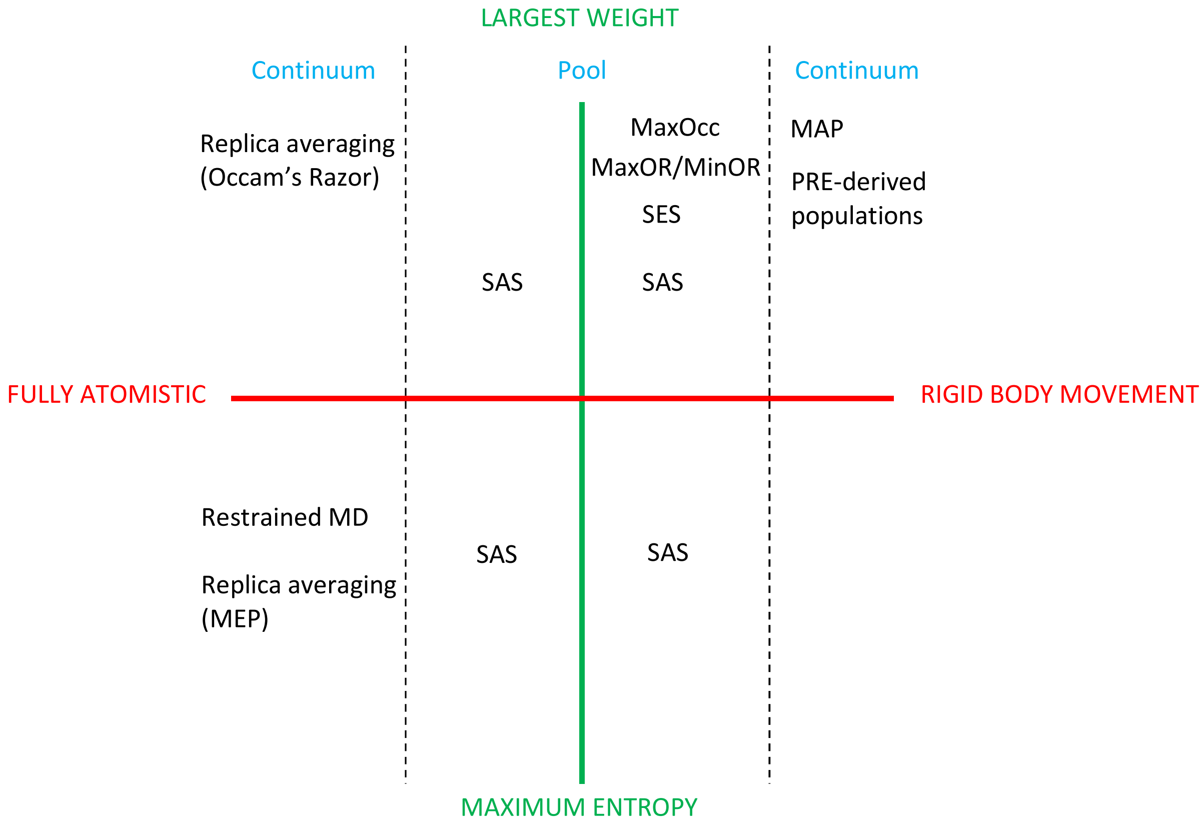


Figure 8. The MaxOcc of a conformation (left panel) is the maximum weight providing a TF larger than a threshold (dotted line) defined depending on the lowest TF. In this panel, the MaxOcc values for three selected conformations are 0.1, 0.3 and 0.45. The MinOR and MaxOR (right panel) are the lowest and largest weight, respectively, of an ensemble of conformations selected within a predefined region of the conformational space which, together with other conformations selected outside this region, provide a TF below the threshold.

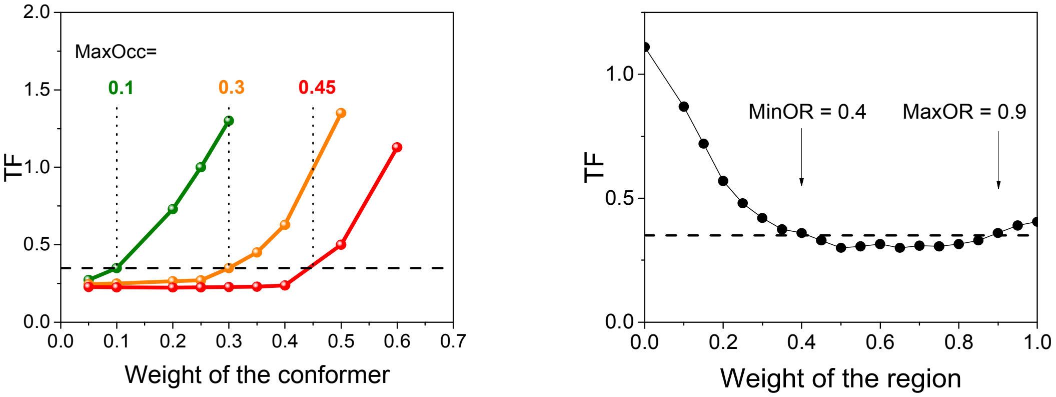
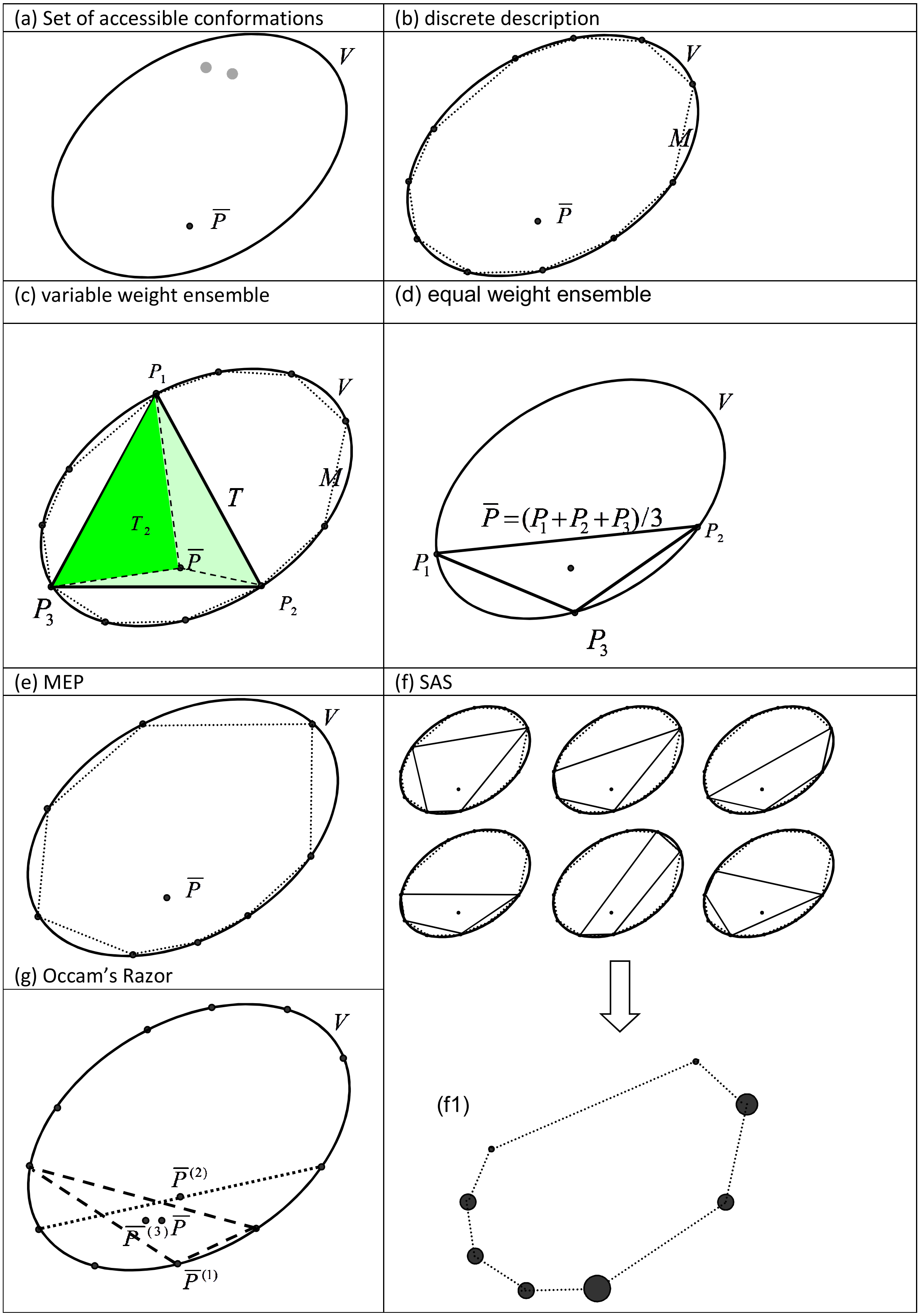


Figure 9. Geometric interpretation of ensemble averaging approaches



(a) Set of accessible conformations

(b) discrete description

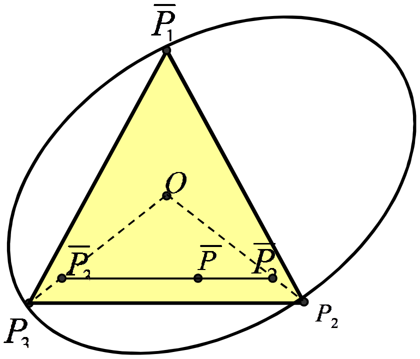
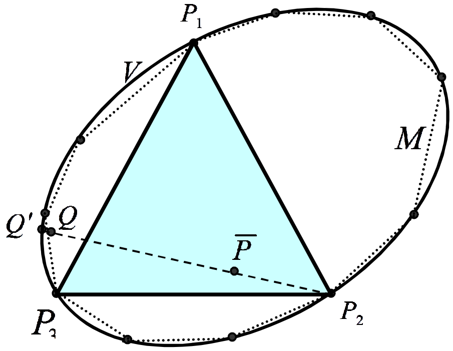
(c) variable weight ensemble

(d) equal weight ensemble

(e) MEP

(f) SAS

(g) Occam's Razor

(h) SES	(i) MAP, MaxOcc
 $\overline{P}_1$ O $\overline{P}_2$ $\overline{P}$ $\overline{P}_3$ P_3 P_2	 P_1 V M Q' Q $\overline{P}$ P_3 P_2

Reference List

(1) L. C. Wang, Y. X. Pang, T. Holder, J. R. Brender, A. V. Kurochkin and E. R. P. Zuiderweg, *Proc.Natl.Acad.Sci.USA*, 2001, ***98***, 7684-7689.

(2) E. Z. Eisenmesser, D. A. Bosco, M. Akke and D. Kern, *Science*, 2002, ***295***, 1520-1523.

(3) K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson and M. Vendruscolo, *Nature*, 2005, ***433***, 128-132.

(4) Y. J. Huang and G. T. Montelione, *Nature*, 2005, ***438***, 36-37.

(5) M. Fragai, C. Luchinat and G. Parigi, *Acc.Chem.Res.*, 2006, ***39***, 909-917.

(6) S. R. Tzeng and C. G. Kalodimos, *Nature*, 2009, ***462***, 368-372.

(7) P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon and M. K. Rosen, *Nature*, 2012, ***483***, 336-340.

(8) Y. E. Shapiro, *Prog.Biophys.Mol.Biol.*, 2013, ***112***, 58-117.

(9) D. A. Torchia, *Prog.Nucl.Magn Reson.Spectrosc.*, 2015, ***84-85***, 14-32.

(10) E. Ravera, L. Salmon, M. Fragai, G. Parigi, H. M. Al-Hashimi and C. Luchinat, *Acc.Chem.Res.*, 2014, ***47***, 3118-3126.

(11) M. Fragai, C. Luchinat, G. Parigi and E. Ravera, *Coord.Chem.Rev.*, 2013, ***257***, 2652-2667.

(12) H. van den Bedem and J. S. Fraser, *Nat.Methods*, 2015, ***12***, 307-318.

(13) Jaynes, E. Where do we stand on maximum entropy?; In *The Maximum Entropy Formalism*; Levine, R., Tribus, M., eds. MIT press: Cambridge,MA, 1979; pp 1-104.

(14) S. Kullback and R. A. Leibler, *Ann.Math.Statist.*, 1951, ***22***, 79-86.

(15) G. M. Clore and C. D. Schwieters, *J.Am.Chem.Soc.*, 2004, ***126***, 2923-2938.

(16) J. Iwahara, C. D. Schwieters and G. M. Clore, *J.Am.Chem.Soc.*, 2004, ***126***, 5879-5896.

(17) M. Longinetti, C. Luchinat, G. Parigi and L. Sgheri, *Inv.Probl.*, 2006, ***22***, 1485-1502.

(18) I. Bertini, Y. K. Gupta, C. Luchinat, G. Parigi, M. Peana, L. Sgheri and J. Yuan, *J.Am.Chem.Soc.*, 2007, ***129***, 12786-12794.

(19) I. Bertini, C. Luchinat, M. Nagulapalli, G. Parigi and E. Ravera, *Phys.Chem.Chem.Phys.*, 2012, ***14***, 9149-9156.

(20) A. N. Volkov, M. Ubbink and N. A. J. Van Nuland, *J.Biomol.NMR*, 2010, ***48***, 225-236.

(21) W. Andralojc, K. Berlin, D. Fushman, C. Luchinat, G. Parigi, E. Ravera and L. Sgheri, *J.Biomol.NMR*, 2015, ***62***, 353-371.

(22) L. Russo, M. Maestre-Martinez, S. Wolff, S. Becker and C. Griesinger, *J.Am.Chem.Soc.*, 2013, ***135***, 17111-17120.

(23) J. R. Tolman and K. Ruan, *Chem.Rev.*, 2006, ***106***, 1720-1736.

(24) Y. E. Ryabov and D. Fushman, *Magn.Reson.Chem.*, 2006, ***44***, S143-S151.

(25) Y. E. Ryabov and D. Fushman, *J.Am.Chem.Soc.*, 2007, ***129***, 3315-3327.

(26) Bryson M., F. Tian, J. H. Prestegard and H. Valafar, *J Magn Reson*, 2008, ***191***, 322-334.

(27) T. Heyduk, *Curr.Opin.Biotech.*, 2002, ***13***, 292-296.

(28) D. Kajihara, R. Abe, I. Iijima, C. Komiyama, M. Sisido and T. Hohsaka, *Nat.Methods*, 2006, ***3***, 923-929.

(29) T. F. Prisner, A. Marko and S. T. Sigurdsson, *J.Magn Reson.*, 2015, ***252***, 187-198.

(30) I. Kaminker, I. Tkach, N. Manukovsky, T. Huber, H. Yagi, G. Otting, M. Bennati and D. Goldfarb, *J.Magn Reson.*, 2013, ***227***, 66-71.

(31) A. Martorana, G. Bellapadrona, A. Feintuch, G. E. Di, S. Aime and D. Goldfarb, *J.Am.Chem.Soc.*, 2014, ***136***, 13458-13465.

(32) M. V. Petoukhov and D. I. Svergun, *Curr.Opin.Struct.Biol.*, 2007, ***17***, 562-571.

(33) J. H. Lakey, *J R Soc Interface*, 2009, ***6***, S567-S573.

(34) K. Berlin, D. P. O'Leary and D. Fushman, *J.Magn Reson.*, 2009, ***201***, 25-33.

(35) M. Zweckstetter and A. Bax, *J.Am.Chem.Soc.*, 2000, ***122***, 3791-3792.

(36) M. Zweckstetter, *Nat.Protoc.*, 2008, ***3***, 679-690.

(37) C. Camilloni and M. Vendruscolo, *J.Phys.Chem.B*, 2015, ***119***, 653-661.

(38) D. I. Svergun, C. Barberato and M. H. J. Koch, *J.Appl.Crystallogr.*, 1995, ***28***, 768-773.

(39) G. Zaccai, *Eur.Biophys.J*, 2012, ***41***, 781-787.

(40) Y. Ryabov, G. M. Clore and C. D. Schwieters, *J Chem.Phys.*, 2012, ***136***, 034108.

(41) G. M. Clore, A. Szabo, A. Bax, L. E. Kay, P. C. Driscoll and A. M. Gronenborn, *J.Am.Chem.Soc.*, 1990, ***112***, 4989-4991.

(42) J. Iwahara and G. M. Clore, *J.Am.Chem.Soc.*, 2010, ***132***, 13346-13356.

(43) R. Brüschweiler, B. Roux, M. Blackledge, C. Griesinger, M. Karplus and R. R. Ernst, *J.Am.Chem.Soc.*, 1992, ***114***, 2289-2302.

(44) J. B. Hall and D. Fushman, *J Biomol.NMR*, 2003, ***27***, 261-275.

(45) R. E. Dale, J. Eisinger and W. E. Blumberg, *Biophys.J*, 1979, ***26***, 161-193.

(46) J. Schilder, W.-M. Liu, P. Kumar, M. Overhand, M. Huber and M. Ubbink, *Phys.Chem.Chem.Phys.*, 2016, ***DOI: 10.1039/c5cp03781f***.

(47) G. Jeschke, *Prog.Nucl.Magn.Reson.Spectrosc.*, 2013, ***72***, 42-60.

(48) Z. Liu, T. M. Casey, M. E. Blackburn, X. Huang, L. Pham, I. M. S. de Vera, J. D. Carter, J. L. Kear-Scott, A. M. Veloro, L. Galiano and G. E. Fanucci, *Phys.Chem.Chem.Phys.*, 2016, ***DOI: 10.1039/C5CP04556H***.

(49) M. A. Stevens, J. E. McKay, J. L. S. Robinson, H. El Mkami, G. M. Smith and D. G. Norman, *Phys.Chem.Chem.Phys.*, 2016, ***DOI: 10.1039/c5cp04753f***.

(50) D. Kurzbach, A. Vanas, A. G. Flamm, N. Tarnoczi, G. Kontaxis, N. Maltar-Strmecki, K. Widder, D. Hinderberger and R. Konrat, *Phys.Chem.Chem.Phys.*, 2016, ***DOI: 10.1039/c5cp04858c***.

(51) D. Fushman, R. Varadan, M. Assfalg and O. Walker, *Progr.NMR Spectrosc.*, 2004, ***44***, 189-214.

(52) K. Berlin, A. Longhini, T. K. Dayie and D. Fushman, *J Biomol.NMR*, 2013, ***57***, 333-352.

(53) O. Walker, R. Varadan and D. Fushman, *J.Magn.Reson.*, 2004, ***168***, 336-345.

(54) G. M. Clore and C. D. Schwieters, *J.Mol.Biol.*, 2006, ***355***, 879-886.

(55) I. Bertini, C. Luchinat and G. Parigi, *Progr.NMR Spectrosc.*, 2002, ***40***, 249-273.

(56) X. Xu, W. Reinle, F. Hannemann, P. V. Konarev, D. I. Svergun, R. Bernhardt and M. Ubbink, *J.Am.Chem.Soc.*, 2008, ***130***, 6395-6403.

(57) X. Xu, P. H. J. Keizers, W. Reinle, F. Hannemann, R. Bernhardt and M. Ubbink, *J.Biomol.NMR*, 2009, ***43***, 247-254.

(58) S. Olsson, D. Ekonomiuk, J. Sgrignani and A. Cavalli, *J.Am.Chem.Soc*, 2015, ***137***, 6270
6278.

(59) A. Cavalli, C. Camilloni and M. Vendruscolo, *J.Chem.Phys.*, 2013, ***138***, 094112.

(60) W. Boomsma, J. Ferkinghoff-Borg and K. Lindorff-Larsen, *PloS Comput.Biol.*, 2014, ***10***, e1003406.

(61) M. Vendruscolo, *Curr.Opin.Struct.Biol.*, 2007, ***17***, 15-20.

(62) X. Salvatella, B. Richter and M. Vendruscolo, *J Biomol.NMR*, 2008, ***40***, 71-81.

(63) W.-Y. Choy and J. D. Forman-Kay, *J.Mol.Biol.*, 2001, ***308***, 1011-1032.

(64) B. Rozycki, Y. C. Kim and G. Hummer, *Structure*, 2011, ***19***, 109-116.

(65) Y. Chen, S. L. Campbell and N. V. Dokholyan, *Biophys.J.*, 2007, ***93***, 2300-2306.

(66) S. Yang and H. M. Al-Hashimi, *J.Phys.Chem.B*, 2015, ***DOI: 10.1021/acs.jpcb.5b03859***.

(67) K. Berlin, C. A. Castañeda, D. Schneidman-Dohovny, A. Sali, A. Nava-Tudela and D. Fushman, *J.Am.Chem.Soc.*, 2013, ***135***, 16595-16609.

(68) J. R. Allison, P. Varnai, C. M. Dobson and M. Vendruscolo, *J.Am.Chem.Soc.*, 2009, ***131***, 18314-18326.

(69) W. Rieping, M. Habeck and M. Nilges, *Science*, 2005, ***309***, 303-306.

(70) C. K. Fisher, A. Huang and C. M. Stultz, *J.Am.Chem.Soc.*, 2010, ***132***, 14919-14927.

(71) S. Olsson, J. Frellsen, W. Boomsma, K. V. Mardia and T. Hamelryck, *Plos ONE*, 2013, ***8***, e79439.

(72) B. Roux and J. Weare, *J.Chem.Phys.*, 2013, ***138***, 084107.

(73) J. W. Pitera and J. D. Chodera, *J.Chem.Theory Comput.*, 2012, ***8***, 3445-3451.

(74) V. Ozenne, F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, Segard S., P. Bernadó, Charavay C. and M. Blackledge, *Bioinformatics*, 2012, ***28***, 1463-1470.

(75) P. Bernadò, E. Mylonas, M. V. Petoukhov, M. Blackledge and D. I. Svergun, *J.Am.Chem.Soc.*, 2007, ***129***, 5656-5664.

(76) L. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen and M. Blackledge, *J.Am.Chem.Soc.*, 2009, ***131***, 17908-17918.

(77) A. T. Frank, A. C. Stelzer, H. M. Al-Hashimi and I. Andricioaei, *Nucleic Acids Res.*, 2009, ***37***, 3670-3679.

(78) M. Pelikan, G. L. Hura and M. Hammel, *Gen.Physiol.Biophys.*, 2009, ***28***, 174-189.

(79) I. Bertini, L. Ferella, C. Luchinat, G. Parigi, M. V. Petoukhov, E. Ravera, A. Rosato and D. I. Svergun, *J.Biomol.NMR*, 2012, ***53***, 271-280.

(80) A. Huang and C. M. Stultz, *PloS Comput.Biol.*, 2008, ***4***, e1000155.

(81) J. A. Marsh and J. D. Forman-Kay, *J.Mol.Biol.*, 2009, ***391***, 359-374.

(82) K. A. Beauchamp, V. S. Pande and R. Das, *Biophys.J.*, 2014, ***106***, 1381-1390.

(83) M. Sanchez-Martinez and R. Crehuet, *Phys.Chem.Chem.Phys.*, 2014, ***16***, 26030-26039.

(84) M. Groth, J. Malicka, C. Czaplewski, S. Oldziej, L. Lankiewicz, W. Wiczk and A. Liwo, *J.Biomol.NMR*, 1999, ***15***, 315-330.

(85) J. Graf, P. H. Nguyen, G. Stock and H. Schwalbe, *J.Am.Chem.Soc.*, 2007, ***129***, 1179-1189.

(86) L. Salmon, G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter and M. Blackledge, *J.Am.Chem.Soc.*, 2010, ***132***, 8407-8418.

(87) J. Huang and S. Grzesiek, *J.Am.Chem.Soc.*, 2010, ***132***, 694-705.

(88) C. A. Castañeda, A. Chaturvedi, C. M. Camara, J. E. Curtis, S. Krueger and D. Fushman, *Phys.Chem.Chem.Phys.*, 2016, ***DOI: 10.1039/c5cp04601g***.

(89) I. Bertini, A. Giachetti, C. Luchinat, G. Parigi, M. V. Petoukhov, R. Pierattelli, E. Ravera and D. I. Svergun, *J.Am.Chem.Soc.*, 2010, ***132***, 13553-13558.

(90) L. Cerofolini, G. B. Fields, M. Fragai, C. F. G. C. Geraldes, C. Luchinat, G. Parigi, E. Ravera, D. I. Svergun and J. M. C. Teixeira, *J.Biol.Chem.*, 2013, ***288***, 30659-30671.

(91) S. Das Gupta, X. Hu, P. H. J. Keizers, W.-M. Liu, C. Luchinat, M. Nagulapalli, M. Overhand, G. Parigi, L. Sgheri and M. Ubbink, *J.Biomol.NMR*, 2011, ***51***, 253-263.

(92) M. Nagulapalli, G. Parigi, J. Yuan, J. Gsponer, S. Deraos, V. V. Bamm, G. Harauz, J. Matsoukas, M. de Planque, I. P. Gerothanassis, M. M. Babu, C. Luchinat and A. G. Tzakos, *Structure*, 2012, ***20***, 522-533.

(93) L. Sgheri, *Inv.Probl.*, 2010, ***26***, 035003-035003-19.

(94) W. Andralojc, C. Luchinat, G. Parigi and E. Ravera, *J.Phys.Chem.B*, 2014, ***118***, 10576-10587.

(95) W. Andralojc, E. Ravera, L. Salmon, G. Parigi, H. M. Al-Hashimi, and C. Luchinat, *Phys.Chem.Chem.Phys.,* 2016, ***DOI: 10.1039/c5cp03993b.***

(96) C. Tang, C. D. Schwieters and G. M. Clore, *Nature*, 2007, ***449***, 1078-1082.

(97) C. D. Mackereth, T. Madl, S. Bonnal, B. Simon, K. Zanier, A. Gasch, V. Rybin, J. Valcárcel and M. Sattler, *Nature*, 2011, ***475***, 408-411.

(98) A. N. Volkov, J. A. R. Worrall, E. Holtzmann and M. Ubbink, *Proc.Natl.Acad.Sci.USA*, 2006, ***103***, 18945-18950.

(99) Q. Bashir, A. N. Volkov, G. M. Ullmann and M. Ubbink, *J.Am.Chem.Soc.*, 2010, ***132***, 241-247.

(100) C. Tang, J. Iwahara and G. M. Clore, *Nature*, 2006, ***444***, 383-386.

(101) N. L. Fawzi, M. Doucleff, J. Y. Suh and G. M. Clore, *Proc.Natl.Acad.Sci.U.S.A.*, 2010, ***107***, 1379-1384.